Deep Reinforcement Learning Based Approach for Multi-Agent Control of Residential Electric Water Heaters for Distribution Load Management

by

Kevin Abraham

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2021

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The push towards decarbonization and electrification of the society is leading to increased electricity demand. Many countries, including Canada, are utilizing non-greenhouse gas (GHG) emitting sources and renewable energy sources (RES) to meet this increasing demand. Many of the RES, however, are intermittent and uncertain, and are non-load following sources of electricity. Technologies supporting demand flexibility are being increasingly used to respond to intermittent changes in RES supply and meet the power grid requirements by modifying the energy consumption patterns of residential loads.

The work presented in this thesis discusses the application of electric water heaters (EWHs) as flexible and controllable loads. EWHs, accounting for a significant portion (44%) of water heaters in the Canadian residential sector, and being the second largest consumer of electricity in the household sector (20%), are becoming a viable source for providing load flexibility.

This thesis presents a multi-agent reinforcement learning (MARL) approach to address the energy management problem of EWHs. Two agents, the residential aggregator agent (RAA)- for EWH control and the utility agent (UA)- to represent the role of a utility, are designed to interact with each other and the (reinforcement learning) environment to maximize their respective rewards. A novel control algorithm using a binning process is employed by the RAA to control operations of certain groups of EWHs. The multi-agent deep deterministic policy gradient (MADDPG) algorithm is implemented for this problem and used in training the RAA and UA to follow the optimal policy.

The proposed EWH energy management approach is tested for consumers in Ontario, New Brunswick and Quebec which have varying consumer tariff rates. The results demonstrate the ability of the proposed RAA and UA to control the behaviour of EWHs via price incentive signals, thus providing benefits for the consumers and the utility.

Acknowledgements

First and foremost, I'd like to thank God Almighty for His presence in my life, especially over the past two years during this MASc journey. He has been ever gracious and merciful, and I know that He will continue watching over and guiding me throughout my professional career.

Secondly, I am forever grateful towards my family and friends for their support and prayers. Their patience and the sacrifices they've made will never cease to inspire me and I hope to continue making them proud in the coming years.

Thanks, especially towards my supervisors, Professor Kankar Bhattacharya and Dr. Steven Wong for their invaluable support and guidance throughout my MASc studies. They have been patient and understanding over the past two years, all the while challenging me to maximize my potential as an engineer. Their professionalism and attention to detail, in particular, has inspired me, and are qualities I hope to emulate throughout my professional career. I would also like to acknowledge Professors Costa Kapsis and Sagar Naik for serving as members of my committee, and for their insightful comments and feedback.

Last but not least, I wish to acknowledge the support of the Natural Sciences and Engineering Research Council (NSERC) for providing the funding necessary to complete my degree. I am also grateful towards the folks at CanmetENERGY, namely Louis-Philippe Proulx, Dr. Armin Salimi and Thierry Lemieux, for the research collaboration.

Dedication

This thesis is dedicated to all my family and friends who have kept me afloat over the past two years.

Table of Contents

List of Figures

List of Tables

List of Acronyms

AI artificial intelligence

DDPG deep deterministic policy gradient

DL deep learning

DLC direct load control

DR demand response

DRL deep reinforcement learning

DSM demand side management

EV electric vehicle

EWH electric water heater

FHMC Final Hourly Marginal Cost

GHG greenhouse gas

HEP Hourly Energy Price

HOEP Hourly Ontario Energy Price

LSTM long short-term memory

MADDPG multi-agent deep deterministic policy gradient

MARL multi-agent reinforcement learning

ML machine learning

MLP multi-layer perceptron

MPC model predictive control

NN neural network

PDE partial differential equation

RAA Residential Aggregator Agent

RES renewable energy sources

RL reinforcement learning

SoC state of charge

TES thermal energy stored

TOU time-of-use

UA Utility Agent

VB virtual battery

Chapter 1

Introduction

1.1 Motivation

The rise in energy demand, as a result of factors such as population growth and urbanization, is leading to an increased need for more energy resource capacity. For instance, in 2050, 66% of the world's population is expected to live in cities, a noticeable increase from 2014, when 54% of the world population was urban [3]. This worldwide urbanization is increasing the strain on already constrained city infrastructure including transport, energy, water supply, air quality and therefore impacting the health and environment [4]. In Canada, from 1990 to 2017, energy consumption grew by 30%, while on the other hand, by 2018, 82% of the generated electricity was from non-greenhouse gas (GHG) emitting sources [5], which resulted in reduction of total GHG emissions from the electricity sector by 46% from the levels in 2000. In 2018, out of the non-GHG emitting sources, hydro and nuclear accounted for 60% and 15% of the total generation, respectively, and renewable energy sources (RES) made up for 7%. The increased share of generation from non-GHG sources outline a clear push towards the further decarbonization of society. This decarbonization is seen in technologies relating to (but not limited to) the electrification of the building and transportation sectors, which can also lead to increased electricity demand. The increasing electricity demand will also need to be met from non-GHG emitting sources such as RES. In Canada, RES generation has increased by 16% between 2010 and 2018 [5], of which, wind and solar have shown the largest growth as depicted in Figure 1.1 [6].

Many RES (including wind and solar) are considered variable sources of electricity because of their lack of availability at certain times due to uncontrollable external factors [7]. These energy resources, at most times, are not load following, meaning that their

energy generation is independent of demand [8]. As a result, without proper complementary technologies/resources set in place, there will be limits to the amount of RES that can be integrated to the electrical grid. Despite these clear limitations and obstacles, significant efforts are being made towards RES integration because of the following reasons [9]:

- Additional generation capacity is needed to meet the increasing demand for electricity, driven by population/economy growth and decarbonization.
- The need to further reduce GHG emissions and other air pollutants from the environment. For instance, despite accounting for less than 7% of total electricity generation, in 2018, the coal sector was responsible for 63% of electricity related GHG emissions in Canada [5]. This shows that even when 82% of the electricity was from non-GHG emitting sources, there is more work to be done in the area of RES integration.
- Reduction of peak demand and electricity costs through the optimal integration of RES and control of customer loads.

In addition to the aforementioned reasons, government policies are encouraging the growth of more sustainable energy generation technologies, such as, RES. According to the International Energy Agency, $12 trillion dollars (CAD) of investments in the electricity sector are expected over the two decades, ending in 2026 [10]. The share of projects related to RES in these investments is expected to be significant. The growth in RES projects can be seen in the increased penetration of wind energy throughout Canada, where the installed capacity rose from 2,349 MW in 2008 to 13,413 MW in 2019 [11], and is projected to increase to 20,000 MW by 2025 [10].

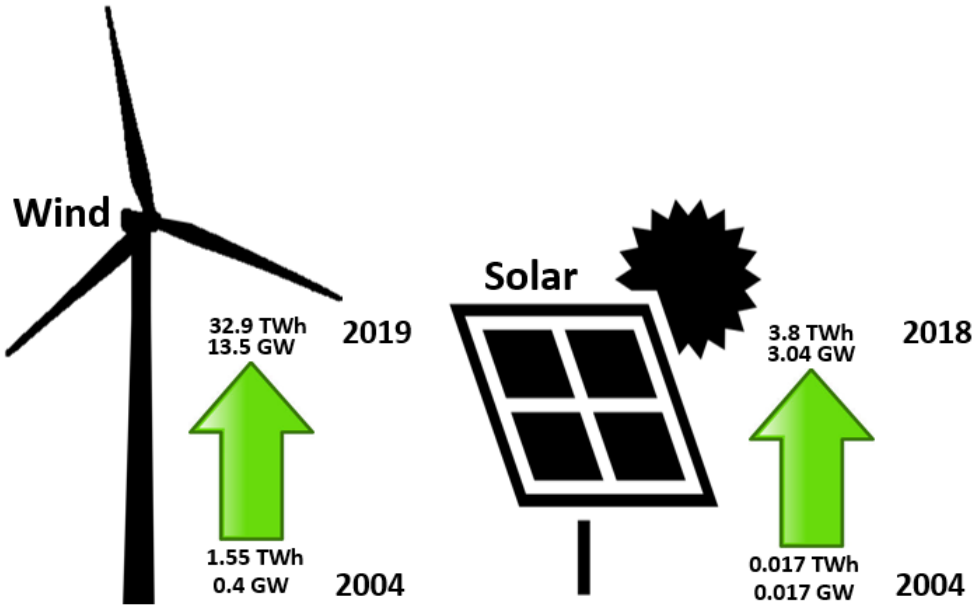


Figure 1.1: Growth of Wind and Solar Generation in Canada

For years, electric utilities and electricity market operators have used traditional demand response (DR) programs and various electricity tariff rates, *i.e.*, time-of-use (TOU) tariffs, to send signals to participating customers to reduce their consumption at certain hours of the day. The new age of smart grids has opened doors to better communication and control technologies which enable "demand flexibility". Demand flexibility (often referred to as demand-side flexibility or load flexibility or flexibility) is a process by which loads (across the commercial, industrial, residential and transportation sectors) can continuously respond to intermittent changes in RES supply, market signals and grid requirements, by modifying their energy consumption patterns [12].

Within the residential sector in particular, thermostatically controlled loads (*e.g.*, electric water heaters (EWHs), air conditioning, space heating/cooling) and electric vehicles (EVs) are some examples of loads capable of providing flexibility. This study will focus on the application of EWHs as flexible and controllable loads to illustrate their potential in flexibility provisions.

Water heaters are commonly used appliances in Canadian households; electricity, natural gas, heating oil and propane are few of the major fuels reported in use for domestic water heating, with electricity and natural gas being the most common [13]. In 2018, domestic water heating was estimated to be the second largest energy end-use in Canadian

households, accounting for approximately 18% of total household energy consumption, exceeded only by space heating [14]. Figure 1.2 illustrates the breakdown of a typical Canadian household electricity consumption, with water heating again being the second largest consumer of electricity at 20%.

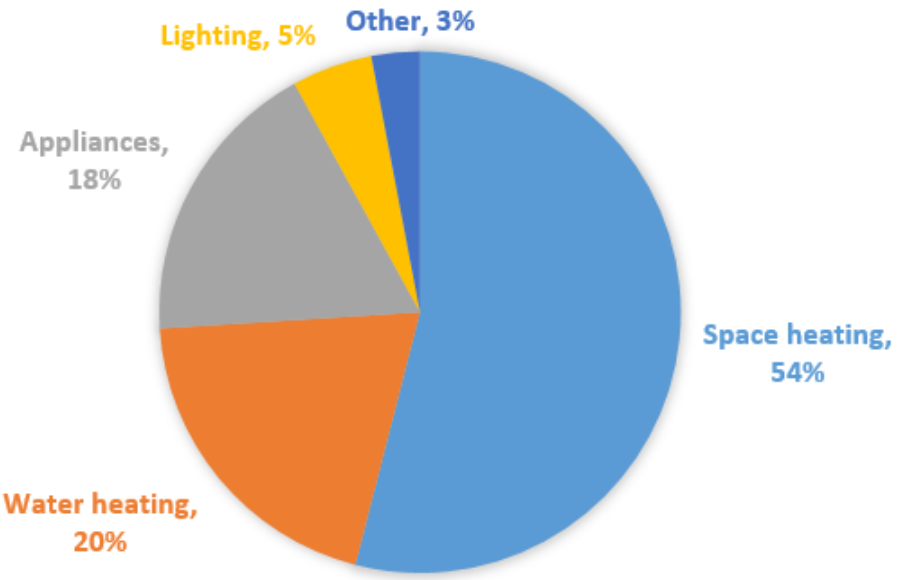


Figure 1.2: Breakdown of Household Electricity Use [1]

In 2018, EWHs accounted for almost 44% of the water heaters in the Canadian residential sector [15]. These EWHs, which have large thermal storage capacities, when aggregated, their control and thermal storage capacity can be leveraged to offer DR, *i.e.*, peak shaving and other flexibility services to the power grid. In addition, consumers can benefit from intelligent control of EWHs through cost savings while maintaining their accustomed levels of comfort [16].

The extensive penetration of RES and the clear transition towards a flexible, bi-directional energy network poses additional challenges to the power grid, such as [17]:

- Integration of advanced communication infrastructure, which results in large volumes of data generated throughout the grid. Proper data storage and processing facilities need to be put in place.

- The presence of highly non-linear EWH loads and uncertainties arising from intermittent RES, and their impact on the power grid.
- Real-time processing of data, autonomous operations and intelligent decision making.

These concerns, and several others, need to be addressed to make the smart grid a reality, and machine learning (ML) based techniques are being investigated for various applications to address these challenges. The ML-based technique, a subset of artificial intelligence (AI), is a data-driven technique, capable of learning patterns and behaviours without relying on predetermined equations and models [17]. In the past, ML has led to significant growth in research in areas such as data mining, communications and medical-imaging. ML applications in the smart grid include price and load forecasting, failure prediction, power generation scheduling, demand side management (DSM), fault detection, and more. Its data-driven nature also means that the ML methods can improve their performance, based on the availability of data. ML techniques excel in tasks relating to adaptive and real-time operation and handling of non-linear systems. Reinforcement learning (RL), one of the sub-fields of ML, is the primary focus of this study and will be further explained in the latter sections. Table 1.1 summarizes the existing problems and limitations in the current power system and benefits of using ML-based techniques [17], [18].

Table 1.1: Motivation behind ML-Based Techniques

Challenges in Smart Grids	**ML Benefits**
Handling of high volumes of data.	Primarily data-driven. Capable of intelligent feature selection and feature extraction for usage of necessary data.
Shift from a non-flexible, unidirectional passive system to a flexible, bi-directional active network.	Capable of performing operations in real-time and can effectively deal with non-linear systems.
High RES penetration in a grid, not originally designed to accommodate them.	Autonomous/intelligent decision making and has the ability to handle complex systems via model-free and model-based techniques. Advances in deep learning in particular, allow applications in wind/solar power control and load forecasting.

1.2 Literature Review

The proper control of common household appliances such as EWHs is a possible option for achieving the benefits associated with RES integration. This section presents a modest review of some of the relevant research, expanding on the discussions from Section 1.1. Research works cited in this section can be divided into two distinct categories: EWH Applications in Grid Services, and DSM through AI approaches.

1.2.1 EWH Applications in Grid Services

This section aims to present key case studies on the implementation of EWH control applications in grid services.

In [19], a heuristic control strategy is presented wherein the heating elements of EWHs are turned off during the pre-defined peak hours (6-10 AM and 4-9 PM) if the water temperature is higher than 50 °C. The study is based on data and patterns from customers in Quebec, Canada, and reports a peak reduction potential of 595 MW (1.68%). The pick-up demand of the EWHs is controlled using a prioritized random activation function spread over one to two hours after the peak period.

In [20], the benefits to the utility are examined through several pilot studies considering bi-directional control of EWHs which allows the utility/aggregator to control EWH behavior, wherein EWHs serve as a flexible energy storage resource. In the maritime region of Canada, the PowerShift Atlantic project tested the aggregation of over 1,000 residential and commercial loads, including EWHs to provide options for energy balancing and ancillary services to facilitate wind energy integration. The Great River Energy, a transmission and generation utility in Minnesota and North Dakota, USA, have been using EWHs for peak load reduction since the 1980s; with over 110,000 participants, positive results have been noted for load-shifting and peak shaving applications.

An ensemble of EWHs are used to provide system frequency regulation services in [21]. A load aggregator assigns each participating EWH a frequency threshold and commits a certain amount of power flexibility for frequency regulation to the grid operator. Depending on the frequency regulation scenario (under-frequency or over-frequency), the committed EWHs either turn on or off to restore the frequency within the assigned range. In case of under-frequency events, this involves reducing the total power consumption (up to the established power flexibility for that control window). The simulations indicate an accurate prediction of EWH behaviour and thereby show potential in providing frequency regulation services to the grid.

In [22], a model for a residential EWH is presented and its response to various DR control signals (both centralized and decentralized) is studied. A 34-bus distribution system with 147 houses with EWHs is used to observe the impact of DR. A centralized signal is sent to the EWHs to aid in frequency regulation of the system by adjusting the EWH temperature settings. The effectiveness of a decentralized signal, which controls the ON/OFF status of the EWH, in emergency situations (*i.e.*, a generator trip), is also studied.

Building on [22], a more accurate EWH model is proposed in [23] which better simulates the dynamic behaviour of EWHs. A partial differential equation (PDE) based model is developed for the EWH and its performance is compared considering over 10 hours worth of field data. To demonstrate the effectiveness of the PDE model, it is compared with that of one-mass (uniform tank water temperature) and two-mass models (water temperature in top half of tank is uniformly close to the setpoint and the temperature in the bottom half is closer to the inlet temperature). The PDE model is able to out-perform other models in certain time steps. The accuracy of the model indicates potential for its role in providing grid services.

The authors in [24] propose an EWH control scheme to ensure consumer comfort amidst the various potential DR solutions. A multi-objective optimization model is used to coordinate thermal comfort and minimize energy consumption, to benefit the homeowner. Three distinct control strategies are examined: (1) maximize energy savings, (2) medium comfort and energy savings and (3) maximize comfort. A week-long simulation is performed in a single person apartment to evaluate the effect of these control strategies. The improvements noted in user comfort and energy consumption efficiency indicate potential for integration of the EWH control scheme with DR solutions.

In [25], the use of EWHs as flexible loads in the Swiss power grid is investigated, with an objective of minimizing the cost of balancing energy. Model predictive control (MPC) is used to communicate with and control the aggregate EWHs. The MPC uses the day-ahead schedule of balancing energy to generate the optimal intra-day DR decisions, which is to control of EWHs during certain times of the day based on real-time measurements and price information.

Though EWH energy consumption is primarily affected by hot water usage, financial incentives/benefits can be provided to consumers to further affect consumption patterns. In [26], a direct load control (DLC) program is proposed and applied which considers a monetary incentive based approach. The DLC method classifies each month of the year as high, medium and low usage months. A pre-defined set of DR hours is established for each classification period and consumers agreeing to the DLC schedule are entitled to monetary benefits. Based on the simulation, the DLC proves to be effective in changing EWH energy

consumption. Results show that high usage months depict higher peak load reduction than medium/low usage months.

Similar to [19] and [20], the methodologies proposed in [27] seek to leverage the aggregated behaviour of EWHs. The potential for peak shifting and frequency response through aggregation of EWHs is examined. The aggregator used in this work is developed based on a C++ program using an open source software framework. The aggregator system uses internet-of-things (IoT) to communicate with multiple EWHs. A communication protocol called CTA-2045 is used so that the aggregator can communicate with various devices supplied by different manufacturers. Unlike the previous works, an EWH emulator was created based on observation from real EWHs, and the emulated EWHs are then connected to the aggregator. The results showed that peak shifting and frequency regulation is attainable using this approach.

It is noted from the review of literature on EWH applications in grid services that many of the works are based on heuristic control approaches which are pre-determined and fixed, and hence are not capable of providing any adaptive or autonomous control features. The control signals in these heuristic approaches are typically of longer duration and are less accommodating to short-term uncertainties. However, due to their simplistic approach, more EWHs can be accommodated in these studies. On the other hand, some works consider intelligent control of EWHs but these lack the scalability aspects – and are not able to consider changing populations of EWHs. When evaluating the flexibility potential of common loads such as EWHs, both intelligent control and scalability aspects need to be accounted for, in order to optimize the grid services.

1.2.2 Demand Side Management: AI Approaches

In [28], a DSM approach is presented for EWH control using neural networks (NNs), with the objective to shift the peaks of the average residential EWH power demand profile to periods of low demand. The DSM involves dividing the EWHs connected to a feeder into groups and controlling their behaviour within a group with different NNs, each group having its own NN. An Elman NN is used to control the EWH behaviour in a given block and it is noted that the proposed approach reduces the peak load.

In [29], the responsiveness of residential consumers' electricity usage under price-driven DR programs is studied. Two models are developed to learn the daily electricity consumption patterns of residential consumers for two types of loads: shiftable/flexible and curtailable. The shiftable loads are from appliances whose usages are restricted to specific

times of the day (*i.e.*, washing machines and dishwashers) while curtailable loads are primarily from devices with adaptable consumption patterns like thermostatically controlled loads. The first model uses a multi-layer perceptron (MLP) architecture which takes the 24-hour ahead electricity price and predicts shiftable load behaviour. The second model is based on a *long short-term memory (LSTM)* network which takes as input the outdoor temperature, power consumption and electricity prices of previous time-steps and outputs the power consumption for the next hour. NNs were the ideal choice for this study, as they were able to learn the electricity consumption dynamics of the households, based on historical data.

The RL approach, a subset of AI, is a type of computational approach built on trial-and-error, dealing with the problem of an agent focused on goal-oriented learning within an environment. In the context of DSM, RL has been applied to tasks related to controlling and scheduling of various components (*e.g.*, domestic appliances, EVs) [30]. The RL methods can be further classified as *tabular* methods, where the low state and action spaces allow for a tabular representation of value functions, and *approximate* methods, used for larger and more complex problems [31]. Since tabular methods are less relevant for the work presented in this thesis, only works relating to approximate methods are discussed below.

In [32], multiple NN architectures via model-free RL algorithms are used to assess the DR potential of a building heat pump. The agent for the RL problem is the heat pump, which aims to maintain the building interior air temperature within established limits. The goal of the agent is to accomplish this task while minimizing the daily electricity cost. The heat pump agent decides whether to draw electric power or remain idle for the duration of the time step. To approximate the state-action value function (Q-function), three different types of NNs are examined: MLP, convolutional NN (CNN) and LSTM. All models performed better than the usual thermostat-controlled implementation, with the MLP implementation being the most favourable due to optimal performance and lowest computation cost.

In [33], RL is used to overcome the stochastic and non-linear dynamics of EWHs and hence develop a controller (agent) to minimize the energy cost. The agent controls the EWH heating element using a binary control signal. The observable state vector consists of current day of the week, time-step, and temperature measurements (based on sensor data) of the tank water. Though the RL agent is able to control the EWH operation, a backup mechanism is put in place to maintain safety and comfort requirements of the consumer. Initially, the EWH is equipped with 50 temperature sensors, with each temperature point used as a value for the state signal for that particular time step. An auto-encoder network is used to reduce the sensory input, thus making the state vector more compact. A fitted

Q-iteration algorithm is then used to approximate the Q-function using a batch of historical data. The proposed method is able to converge to optimal policies and also reduce the energy consumption cost.

In [34], RL is applied to a smart energy hub (SEH) framework to improve the energy efficiency and reduce residential consumer energy costs. The SEH is an upgraded model of the conventional energy hub, optimizing the operations of a residential consumer equipped with combined heat and power, energy storage, auxiliary boiler and heat storage. With the objective of minimizing energy cost, RL is used to control the electricity and natural gas consumption of the aforementioned loads. The results indicate that residential consumers achieve cost savings and reductions in peak load. A new cloud computing system configuration is used in conjunction with the RL implementation to achieve further computational efficiencies.

In contrast to the previously discussed works detailing single-agent methods, multi-agent reinforcement learning (MARL) methods and implementations are gaining momentum as viable solutions in smart grid related problems. The authors in [35] addresses the load frequency control problem via an MARL decentralised approach, by controlling the system frequency at the generator primary and secondary levels. A multi-agent deep deterministic policy gradient (MADDPG) architecture is employed, in which each generator within the network is modelled as an agent. Frequency restoration simulations in a two, four, and eight generator/agent network achieved positive results and indicated potential for applications to larger systems.

Multi-agent systems are investigated in [36], in which the objective is to assess demand flexibility of a commercial building, while maintaining comfort. The environment for this RL problem comprises seven commercial buildings and an aggregation of 27 residential loads. Three agents are described: distribution agent (DA), aggregator agent (AA) and building energy management system (BEMS). The DA monitors the distribution network and generates flexibility requests to be sent to the buildings. The BEMS handles the normal building operations and determines its flexibility potential. The AA acts as a mediator to meet the flexibility demand in an economic and efficient manner. Different algorithms are employed to capture both cooperative and non-cooperative behaviour among the agents.

High penetration of EVs can significantly increase the demand, and thus proper DR methods are needed to ensure that the available capacity is not exceeded. A multi-agent planning and optimal scheduling algorithm for EV charging is described in [37] which uses a collaborative parallel Monte-Carlo tree search to resolve charging conflicts between EVs and optimize the final consumption patterns. An argumentation-based negotiation approach is used to allow the different EV agents to interact/argue with each other to achieve the

optimal proposal. The agents' ability to argue allows them to gain more information about the environment as well as fellow agents, thus leading to a higher proposal acceptance rate. The proposed was noted to flatten the load profile while ensuring that EVs were at a suitable battery state of charge (SoC) level.

Lastly, [38] provides an MARL based approach for home energy management. This data-driven approach is based on the Q-learning algorithm. The multi-agent approach addresses the issue of having different types of loads in a residential house (*e.g.*, non-shiftable loads, EV charging loads, time-shiftable loads, and power shiftable loads). A finite Markov decision process approach is used to model the hour-ahead energy consumption scheduling problem. Feed forward NNs are used to accurately predict the future electricity prices and solar generation patterns. The predictions are provided to the Q-learning algorithm and an optimal energy consumption schedule for different household appliances is obtained. The results show a reduction in electricity costs and improvement in computational efficiency. The control of multiple residential loads to modify energy consumption patterns indicate potential for DR services (*i.e.*, through utility interaction with multiple homes).

The developed models discussed in this sub-section, although effective and yielding positive results, were primarily tested on constrained and non-generic environments. DSM studies need be carried out in real-world environments with seasonal demand variations, differing electricity pricing structures and diverse flexible load characteristics.

1.3 Research Objectives

The review of literature presented in the previous section illustrates the myriad of works considering EWHs and AI based approaches for DSM. The review discussed some novel research on grid services using EWHs and the increasing applications of AI algorithms in the smart grid environment.

The research presented in this thesis uses a fine-tuned granular control of EWHs (minute-based signals), and through the use of RL data driven approach, sends autonomous signals based on DSM requirements and EWH operation characteristics (*e.g.*, hot water draw profiles, and tank water temperature). In addition, to evaluate the overall generalizability of the proposed solution, the models are tested in different regions with seasonal load variations and varying pricing structures.

Based on the aforementioned discussions, the objectives of the research presented in this thesis are as follows:

- Model the aggregated population of EWHs and determine how their behaviour can be leveraged, through a novel binning algorithm, to provide services which benefit the residential consumer and utility.
- Develop NN models to formulate the RL problem for energy management of EWHs as flexible loads. Formulate the necessary RL agents to model the utility and EWH aggregator so as to maximize the objectives of each agent.
- Utilize the deep reinforcement learning (DRL) algorithm, MADDPG, to simulate the interactions and behaviours of the two agents (EWH aggregator and utility) within the RL environment.
- Test the behaviour of the trained agents on real data to simulate the multi-agent operation on a residential feeder.
- Gain insight into how the developed algorithm operates in different Canadian geographic regions with different electricity pricing structures. The regions examined in this study are the Canadian provinces of Ontario, New Brunswick and Quebec.

1.4 Outline of the Thesis

The thesis is structured as follows: *Chapter 2* provides a background to the main concepts addressed in this thesis, such as the model used to simulate EWH operation, distribution system operation and DSM. The theoretical background to RL, MARL, deep deterministic policy gradient (DDPG) and MADDPG are also presented. *Chapter 3* discusses the model architecture of the proposed MARL problem and the environment and its dynamics. In addition, the state, action, reward signals and objectives of each agent are presented. *Chapter 4* presents the detailed results and discussions for the case studies considering the proposed models and are tested on varying pricing structures from several jurisdictions in Canada such as Ontario, New Brunswick and Quebec. *Chapter 5* presents a summary and the main contributions of the research and outlines areas for future research.

Chapter 2

Background

This chapter reviews the theoretical background of the main concepts pertaining to the work presented in this thesis. Firstly, the operational dynamics of EWHs are discussed, such as the generation of draw, load and temperature profiles. Secondly, the RL background as well as the mechanics behind the MADDPG algorithm are reviewed in great detail. Tables 2.1 and 2.2 list the paramaters and variables used in Section 2.1 for the creation of the EWH draw, load and temperature profiles.

Table 2.1: EWH Variables

Variables	Description
$\theta_n(t)$	Water heater temperature at min t ($^{\circ}C$)
a	Fraction standby temperature drop (min^{-1})
A	Temperature drop from hot water drawn ($\frac{^{\circ}C}{min}$)
DEE	Daily energy extraction (MJ)
R	Rate of temperature gain when charging ($\frac{^{\circ}C}{min}$)
α_0	Probability of transitioning from no hot water draw to water draw
α_1	Probability of transitioning from hot water draw to no water draw

Table 2.2: Parameters

Parameters	Value	Description
θ_a	22	Ambient temperature ($^\circ C$)
$\theta_{wh}^{in,w}$	1.5	(Winter) EWH input water temperature ($^\circ C$)
$\theta_{wh}^{in,s}$	23	(Summer) EWH input water temperature ($^\circ C$)
θ_{wh}^{max}	61	EWH maximum water temperature ($^\circ C$)
θ_{wh}^{min}	55	EWH minimum water temperature ($^\circ C$)
θ_{wh}^{set}	57.5	EWH setpoint water temperature ($^\circ C$)
c	4.184	Specific heat capacity of water ($\frac{kJ}{kg \cdot ^\circ C}$)
e	6.4	Extraction rate ($\frac{L}{min}$)
$elem(t)$	0 or 1	EWH element operation at min t
m	270	Mass of tank water (kg)
P_{rated}	4.5	EWH element rating (kW)
$q(t)$	0 or 1	State of hot water drawn at min t
v_{QC}	266.34	Daily Quebec hot water demand (L)
v_{ON}	175.5	Daily Ontario hot water demand (L)
v_{NB}	283.68	Daily New Brunswick hot water demand (L)
w	70.7	Thermal conductance ($\frac{J}{min \cdot K}$)

2.1 Electric Water Heater

This section provides the theoretical background relating to the dynamics of EWH operation. The types of residential EWHs available in the market today vary based on consumer requirements and preferences. A few of these factors include capacity (180 l and 270 l being the most common) and rated power of EWH heating element (3.0 kW and 4.5 kW being the most common). The number of heating elements typically range from one to three, with two being the most common [39]. Additional components present in most EWHs are represented in Figure 2.1. The positions of these components shown in Figure 2.1 may vary across different EWHs.

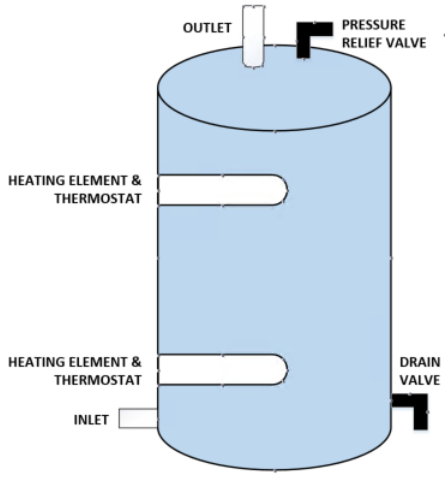


Figure 2.1: Components in a Typical EWH

2.1.1 General Operation

In a water heater, cold water enters through the inlet valve and hot water exits through the outlet valve. Once hot water is used at a faucet, cold water quickly fills the bottom of the tank, thereby maintaining water heater capacity at all times.

A thermostat is used to measure the water temperature inside the EWHs. The temperature set point of EWHs has a recommended range between 50 °C and 60 °C (depending on manufacturer specifications) [39], which accounts for important factors such as preventing legionella growth (a type of bacteria known to cause a rare strain of pneumonia), potential hot water scalding, water heater component degradation and optimal energy consumption [39], [40]. A temperature dead-band, typically within ±5°C, is used to determine the on/off behaviour of the heating elements and to ensure that hot water is available at all times.

In the case of EWHs with two heating elements, the heating elements are interlocked to prevent simultaneous operation, with the priority given to the top element. For instance, once the tank is filled with water, the top element turns on to heat up the water near the top of the tank. Once the water in the upper tank is heated to the desired temperature, the top element is turned off and bottom element is turned on to similarly heat the water near the bottom of the tank. This operation leads to a stratified temperature profile of the water in the EWH, with the hottest water being near the top of the tank. An example of a possible temperature stratification profile is shown in Figure 2.2. Thus, hot water draw patterns and inlet water temperatures have an impact on the water temperature

stratification profiles in EWHs.

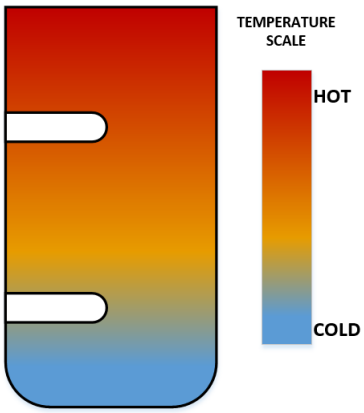


Figure 2.2: EWH Temperature Stratification

Figure 2.3 shows the simplified EWH representation used in this research [39], which uses only one heating element and assumes a uniform water temperature throughout the tank (no temperature stratification). These assumptions are made to simulate the power draw behaviour of one tank in a population hundreds or thousands, and is not inhibited by the simplified model [39]. Parameters and variables of the EWH used in this study are given in Tables 2.1 and 2.2.

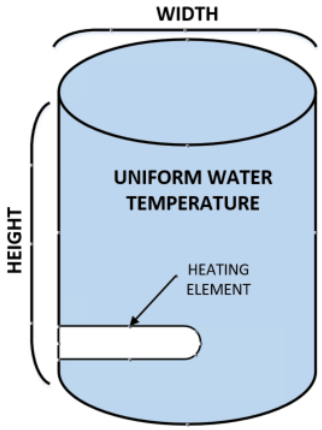


Figure 2.3: Simplified EWH Representation

2.1.2 Hot Water Draw Profiles

Obtaining hot water draw profiles of EWHs is a critical component of simulating a population of these EWHs and are necessary to calculate the hot water temperature at a given time. A two-state Markov chain [39] is used to obtain the water draw states of the EWH, as shown in Figure 2.4.

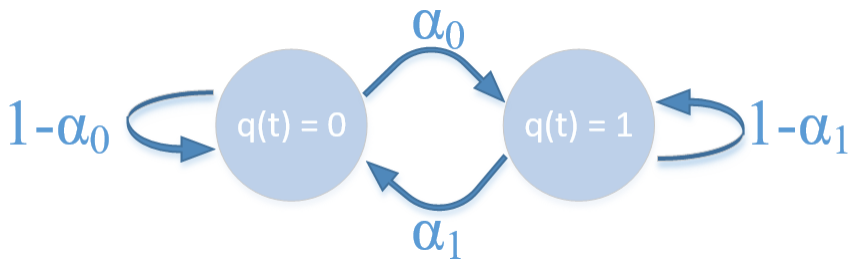


Figure 2.4: Two-State Markov Chain for EWH Water Draw States

where

$$q(t) = \begin{cases} 0, & \text{if hot water not drawn (per minute)} \\ 1, & \text{if hot water drawn (per minute)} \end{cases} \tag{2.1}$$

α_0 : Probability of transitioning from no hot water draw to hot water draw (2.2)

α_1 : Probability of transitioning from hot water draw to no hot water draw (2.3)

The values of α_0 and α_1 used for this study and the algorithm used to obtain values for *q(t)* is taken from [39]. Figure 2.5 provides the draw profile of multiple populations of EWHs.

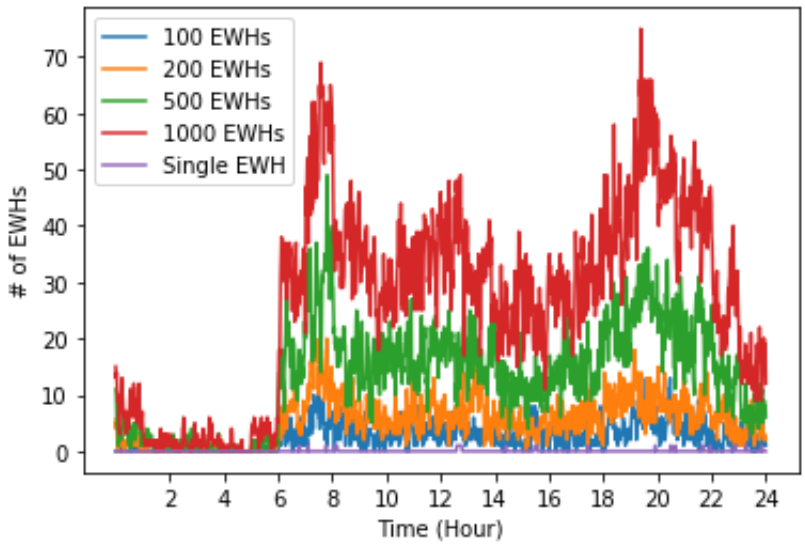


Figure 2.5: Hot Water Draw Profiles for Different Populations of EWHs

Figure 2.5 depicts a simulated profile of the water drawn by a single, and groups of 100, 200, 500 and 1,000 EWHs over one day (*i.e.*, 1,440 minutes), obtained using the 2-state Markov model discussed earlier. It is noted that the profiles of the aggregated EWHs follow a similar pattern – high usage during the morning and evening hours. The research presented in this study outlines the scalability feature of the algorithm: the similarities in usage patterns of large populations of EWHs allow using small populations during the training phase of the RL implementation, which can ease the computational burden. The trained models can then be applied to large EWH populations. The RL component of the problem is discussed later in this chapter. Additionally, as noted from the profiles of hot water drawn (Fig. 2.5), there is clearly a diversified usage by EWH consumers. For instance, in the simulation of 1,000 EWHs for any given minute on this particular day, a maximum of 80 consumers draw hot water. This diversified behaviour is important and is further discussed in Section 2.1.3.

2.1.3 Generating Temperature and Load Profiles

The EWH heating elements convert electrical energy into thermal energy, which is then absorbed by the water mass. It uses the thermostat and heating elements to maintain the water temperature within set limits (established by the temperature dead-band). The

water temperature is affected by three components: standby heat loss, state of heating element and hot water demand.

Standby heat losses, in the context of EWHs, refer to the losses associated with maintaining the tank water temperature during idle times (heating element OFF and no hot water demand) [41]. Standby losses are a function of the ambient and tank water temperature and tank insulation, and is calculated as follows:

$$a = \frac{w}{1000 \cdot c \cdot m} \tag{2.4}$$

The Daily Energy Extraction (DEE) in MJ, from the EWH by virtue of hot water drawn, varies across different households, and especially across different regions [39]. This is primarily due to the difference in hot water demand between different geographical regions caused by factors such as weather conditions. The DEE is calculated as follows:

$$DEE = v \cdot c \cdot \frac{\theta_{wh}^{set} - \theta_{wh}^{in}}{1000} \tag{2.5}$$

As mentioned earlier, the EWH heating element and thermostat device works together to maintain the tank water temperature within the dead-band. Rate of temperature gain ($^{\circ}C/min$) when EWH heating element is turned ON is calculated as follows:

$$R = \frac{60 \cdot P_{rated}}{m \cdot c} \tag{2.6}$$

Temperature change of the water in EWHs is caused by standby heat losses and hot water usage. Since ambient temperature in residential households remains fairly unchanged throughout the year and tank insulation is generally good, standby losses are generally constant and low [39]. Hot water usage is therefore the main contributor to temperature drop in an EWH. Cold water replaces the hot water when hot water is drawn, and thus, peak hot water usage times can result in a significant reduction in average temperature. Temperature drop ($^{\circ}C/min$) from hot water drawn is calculated as follows:

$$A = \frac{1000 \cdot DEE \cdot e}{v \cdot m \cdot c} \tag{2.7}$$

The state of the EWH heating element at time *t*, *elem(t)*, is determined from the average tank water temperature at time *t-1* using the following process:

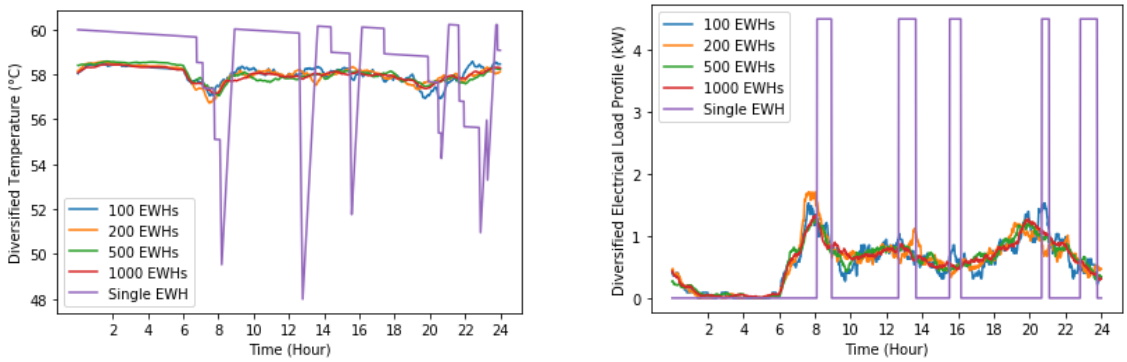
- If tank water temperature is below the minimum allowed, *i.e.*, $(\theta_n(t-1) < \theta_{wh}^{min})$, EWH heating element is activated/remains activated; *elem(t)* = 1.
- If tank water temperature is above the maximum allowed, *i.e.*, $(\theta_n(t-1) > \theta_{wh}^{max})$, EWH heating element is deactivated/remains deactivated; *elem(t)* = 0.
- Else EWH heating element continues operation in prior state; *elem(t)* = *elem(t-1)*.

Using equations 2.4 - 2.7, the temperature change dynamics models the average water temperature of an EWH, as given below.

$$\Delta\theta_n = -a(\theta_n(t) - \theta_a) - A(q(t)) + R(elem(t)) \tag{2.8}$$

The first component of (2.8) represents the temperature change due to standby heat losses, the second component represents the temperature drop associated with hot water usage (the formulation of hot water usage profile is discussed in Section 2.1.2), and the last component denotes the temperature gain when the EWH heating element is ON.

Using the aforementioned temperature change function, the temperature and load profiles of EWHs are simulated for one day or 1,440 minutes, as shown in Figure 2.6. Certain differences can easily be noted when comparing the temperature and load profiles of an individual EWH with the average profiles of 100, 200, 500, and 1,000 EWHs, which represent their diversified behaviour; the average temperature profile remains fairly steady between 57°C and 58 °C while for an individual EWH, the fluctuation is much more pronounced. The same is true for the load profile of a single EWH, in which the heating element activation doesn't seem to follow any specific pattern. Individual houses may have varying usage patterns, hence the increased variations for the single EWH profile.



(a) Average Temperature Profiles

(b) Average Electric Load Profiles

Figure 2.6: Water Temperature and Load Profiles of Different Populations of EWHs

Certain patterns can be observed from the average profiles of the larger population of EWHs. For instance, Figure 2.6 shows a decrease in average water temperature and a corresponding increase in average electrical demand of the households at certain times of the day (between 6-8, and 19-21 hours), which can be attributed to hot water use at these times. Peak times of hot water usage typically occur in the mornings and evenings and this reduces the average tank temperature, causing the EWHs to re-charge by turning ON the heating element, thus increasing the electrical demand. Lastly, it is important to note that Figures 2.5 and 2.6 were obtained using an inlet water temperature of 1.5 $^{\circ}C$. Increase in inlet water temperature will lead to reduction in consumed energy, thus potentially affecting the average load profiles.

Information on the diverse behaviour of loads, obtained from the average load profiles, is necessary because loads do not normally peak at the same time. Consideration of load diversity is crucial for distribution system planning, and is also required for equipment sizing. In the same context, understanding and leveraging the aggregated/diversified behaviour of certain household loads (EWHs in this study) can help with DR schemes [12]. Based on the above analysis, this thesis will consider the behaviour of 100 EWHs to emulate the behaviour of large populations of EWHs. This assumption is necessary as the RL training component requires an accurate representation of the behaviour of EWH populations. Simulating large populations of EWHs during training is a significant computational burden, and thus 100 EWHs is a sufficient minimum.

2.2 Machine Learning

ML is a sub-field of AI, wherein ML algorithms find patterns in data sets through one of four learning paradigms: supervised, unsupervised, semi-supervised and RL. This thesis focuses on the application of a DRL algorithm to simulate the interactions and behaviours of EWH aggregator and utility. This section focuses on principles relating to RL, DRL and the relevant algorithm used for this thesis - the MADDPG algorithm. Tables 2.3, 2.4 and 2.5 present the variables, parameters and sets used in Section 2.2.

Table 2.3: RL Variables

Nomenclature	Description
a	Action signal
a'	(next) Action signal
d	Terminal state of episode
$Q(s,a)$	Q-value
r	Reward signal
s	State signal
s'	(next) State signal
y	Target Q-value

Table 2.4: NN Parameters

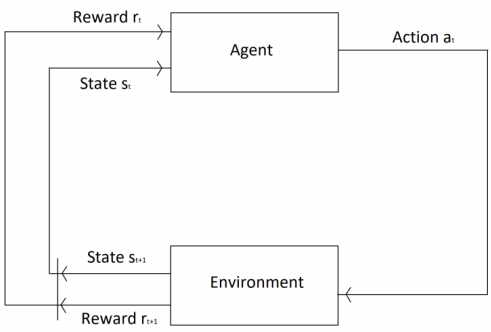
Nomenclature	Description
π	Policy
γ	Discount factor
τ	Update parameter ($\tau << 1$)
Q	Critic network
μ	Target network
Q'	Target critic network
μ'	Target actor network
θ^Q	Critic network parameters
θ^μ	Actor network parameters
$\theta^{Q'}$	Target critic network parameters
$\theta^{\mu'}$	Target actor network parameters

Table 2.5: RL Sets

Nomenclature	Description
A	Set of actions for all agents
B	Sample batch from D ($B = (s, a, r, s', d)$)
D	Replay buffer
R	Set of rewards for all agents
S	Set of states for all agents
S'	Set of next states for all agents

2.2.1 Reinforcement Learning

RL uses a trial-and-error approach to make optimal decisions for a given objective; it has been used to solve many problems and is being considered as a viable approach to solve many decision and control problems in power systems [42]. In RL, an agent, typically representing the component of the system required to learn and act in an optimal manner, takes actions (a) in an environment in order to maximize its reward (r). Depending on the action taken by the agent, the environment changes, which is governed by the environment dynamics. The environment is represented by a state signal (s). The agent eventually learns the optimal policy (π) through its interactions with the environment and the resulting rewards (Figure 2.7 (a)). Figure 2.7(b) presents a generic pseudocode for the RL problem.



(a) RL Operation

Initialize Agent
while *Episode e <Emax* **do**
  Initialize Environment
  for *Time* $t = 1$ ***to*** T **do**
    Observe state s_t
    Select action a $\in A$
    Observe reward r and new state s$'$
    Update policy π
    Transition to new state s_{t+1}
  end
end

(b) Generic RL Pseudocode

Figure 2.7: Basic features of RL Algorithm [2]

The duration of an RL simulation is determined from the initial and terminal states of the problem, also known as an *episode*. The goal of the agent is to maximize the total reward it receives during an episode. It is important to note that the episodes are independent of each other [43]. To properly train an agent in an RL problem, the agent needs to interact with the environment over multiple episodes.

In problems with a termination state, the cumulative reward is known as the *return* G. A discount factor (γ) in the range of [0, 1] is necessary for the agent to prioritize short-term rewards over long-term rewards; the total return over a full episode is given by [31]:

$$G_\pi(t) = \sum_{t=t'}^{T} \gamma^{t-t'} \cdot r(s_t, s_{t+1}, a_t|\pi) \tag{2.9}$$

The term $r(s_t, s_{t+1}, a_t|\pi_t)$ refers to the reward received while transitioning from state s_t to s_{t+1} by taking action a_t, which is chosen based on policy π_t. The reward over a full episode can be calculated by setting t' to 0.

2.2.2 Deep Reinforcement Learning

DRL is a subset of ML that utilizes components present in deep learning (DL) and RL. DL is a form of ML that uses NNs to transform a set of inputs into a set of outputs. NNs are widely used for nonlinear function approximation [31]. Figure 2.8 shows a generic feed forward NN, meaning a NN with no loops in the network.

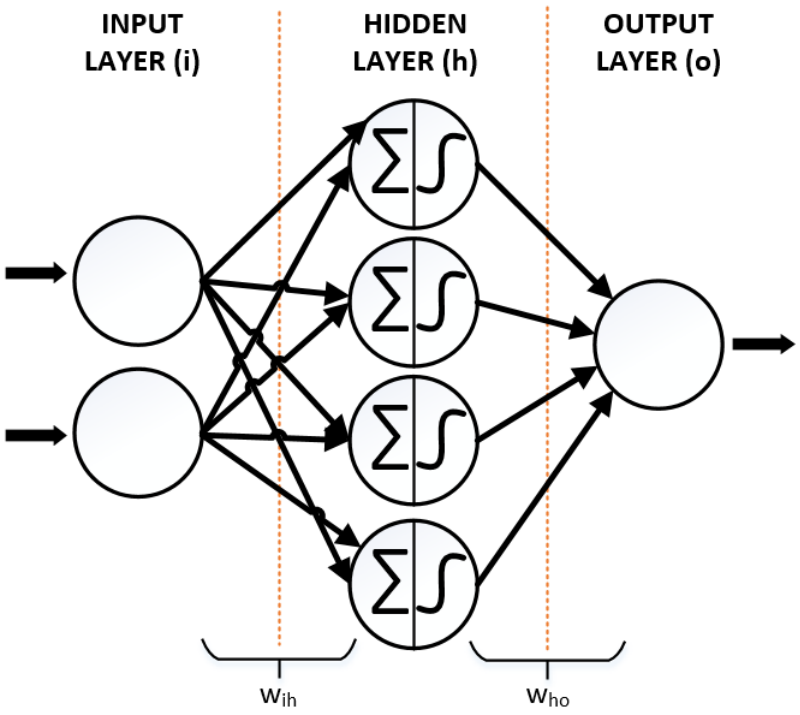


Figure 2.8: Depiction of a Basic Neural Network

The NN in Figure 2.8 includes a two-neuron input layer, four-neuron hidden layer, and one-neuron output layer. Typically, these neurons compute a weighted sum of their input signals and apply the resulting value to a nonlinear function, called the activation function [31]. This produces the output of the particular neuron. Commonly used activation functions include *rectified linear* (ReLU), *sigmoid* and *tanh* functions [44]. It is also important to note that different layers within the same NN may use different activation functions.

Approximating complex functions, as in many AI problems, require deeper NNs, meaning more hidden layers. These additional hidden layers are used to extract better features than shallower NNs, and thus can learn more effectively. In use of NNs to classify images, the first layer may train itself to recognize simple patterns like edges, the second layer may train itself to recognize composites of edges such as shapes, and so on. This training is carried out by adjusting the weights of each neuron link in a direction aimed at improving the NN's overall performance, as established by an objective function.

In several real-world problems, the state and action space of the agent can be quite large-dimensional and thus cannot be solved using traditional RL algorithms. For example,

the application of RL in a simple tic-tac-toe game is explained in [31]; the game takes place in a conventional 3x3 grid, where one player plays "X", the other plays "O", and the winner is one who places three of its marks in a row, first. Using a tabular RL method, a table of numbers is created for each slot of the 3x3 grid which represents the probability (or the state's value) of winning the game from that state, i.e., if the value of state 1 is greater than that of state 2, the player has a higher chance of winning from state 1. The tic-tac-toe game presents a very small and finite state set (3x3 grid); and the action set in the game is also finite, as in each turn the agent is required to mark one of the available grid spots. However, many real-world applications (robotics being a common field for RL study) require continuous state/action spaces defined by means of continuous variables. In the case of robotics applications, a state space could be defined by continuous variables like position, torque, velocity and an action space set by continuous variables like the angle values (which govern the torques on the joints of a robotic arm). The usual approach has been to discretize the continuous variables. However, this quickly leads to a combinatorial explosion, and thus the well known "curse of dimensionality" [45].

NNs used in DL are function approximators and can be useful in such RL problems. NNs provide the agent the ability to generalize based on its past experience (*i.e.*, the states and actions from previous time steps) rather than discretizing the continuous variables. In this manner, when the agent encounters a new state, it is able to select an action based on its previous encounters with similar states. Incorporating DL methodologies in RL algorithms has proven successful in solving complex problems [46]. In order to maximize return, many approaches are available, and the specific method to use is, to an extent, problem dependent. For the problem addressed in this thesis, a DRL algorithm, specifically the DDPG algorithm, will be employed, which is discussed in sections 2.2.3 and 2.2.4.

2.2.3 Deep Deterministic Policy Gradient

DDPG is a commonly used algorithm in RL problems with complex environments (a distribution grid, for example) and continuous action spaces. The DDPG algorithm uses *policy gradient* (PG) based methods and actor-critic network architecture, both of which are known to be useful in attaining convergence in large continuous state and action spaces [47]. The objective of an RL agent is to maximize the expected reward (2.11) when following a particular policy π. Deep NNs are used to approximate the agent's policy by taking observations from the environment as input, and outputting actions based on said policy. If θ^π denotes the policy parameters and p the performance of the policy, the policy gradient approach updates θ^π approximately using the gradient approach, given by [48]:

$$\Delta\theta^{\pi} \approx \alpha \cdot \frac{\partial p}{\partial \theta^{\pi}} \tag{2.10}$$

where α is a small positive step size. This gradient approach requires a large sample size to arrive at an optimal policy.

The actor network, also denoted as "μ" network, learns an approximation of the optimal behaviour policy, which is a mapping from the state space to action space. The critic network, also denoted as "Q" network, as indicated in its name, is used to critique the actions of the actor network. It accepts the states and actions as inputs and learns an approximate action-value function which guides the training of the actor network. The interaction between the actor and critic networks is illustrated in Figure 2.9:

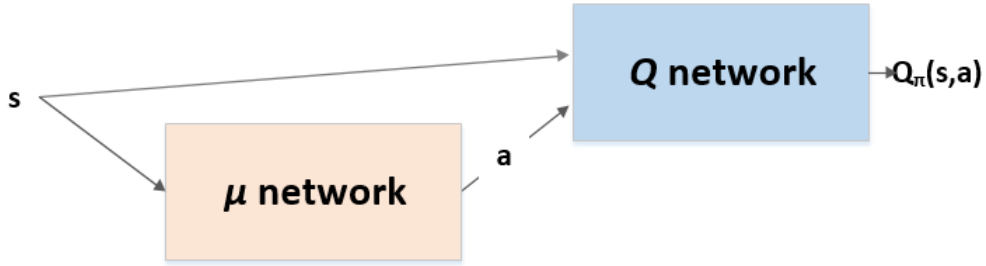


Figure 2.9: Interaction Between Actor and Critic Networks

The action-value function, also known as the Q-value or Q-function, denoted by $Q_{\pi}(s, a)$, represents the expected return from an action in a specific state, while following policy π, and is given as follows [47]:

$$Q_{\pi}(s, a) = E_{\pi}[G|s, a] \tag{2.11}$$

DDPG utilizes the Bellman equation to obtain the optimal Q-function ($Q^*(s, a)$) for a given state. The expected return from starting at state s, taking action a under the optimal policy, is equal to the immediate reward plus the maximum of expected discounted return from the next state and action.

$$Q^*(s, a) = E_{\pi}[R_{t+1} + \gamma \cdot \underset{\mathbf{a'}}{max}\, Q^*(s', a')] \tag{2.12}$$

In addition to the actor and critic networks, DDPG utilizes two more networks; the target actor ("μ'") and target critic ("Q'") networks. From (2.12), it is noted that $Q^*(s, a)$

is calculated from $Q^*(s', a')$. However, since s and s' are only one step apart, they may be similar and NNs may struggle to distinguish between them. Also, while the NN parameters are being updated to make $Q(s, a)$ closer to the desired result, $Q(s', a')$ may get altered. This "moving target" nature of the problem makes training quite unstable. Target networks are used to save a copy of the NN (thereby storing the trained NN parameters) and obtaining $Q(s', a')$. Thus, results from the target networks (μ' and Q') can be used to train the main actor and critic networks. Target networks are used to add stability to training, which is accomplished through slow target network updates (also known as soft updates) [49]. Utilizing DDPG, each agent is modelled by four networks: actor network, critic network, target actor network and target critic network.

In DDPG, *experiences* are stored in the (replay) buffer, typically in the form of a tuple, consisting of state, action, reward and next state values (s_t, a_t, r_t, s_{t+1}). The replay buffer is simply a set of previous experiences obtained over numerous time steps and episodes as the agent interacts with the environment. At each time step, the NNs are updated by uniformly sampling a mini-batch from the buffer [47]. An adequately large replay buffer will contain a wide range of experiences. A small and less diverse replay buffer is detrimental to training the agent, as the access to limited information may lead to overfitting.

In order for the agent to obtain optimal performance, the actor and critic network parameters have to be updated accordingly.

Actor Update: The actor network update process is rather straightforward compared to that of the critic network. The actor loss is simply the sum of Q-values for the states. For computing the Q values, the critic network is used to pass the action computed by the actor network. To maximize returns, a gradient ascent method is employed as follows [47]:

$$\begin{aligned} J_\mu &= \frac{1}{|B|} \sum_B Q(s, \mu(s|\theta^\mu)) \\ \nabla_{\theta^\mu} J(\theta) &= \frac{1}{|B|} \sum_B \nabla_\mu Q(s, \mu(s|\theta^\mu)|\theta^Q) \nabla_{\theta^\mu} \mu(s|\theta^\mu) \end{aligned} \tag{2.13}$$

Critic Update: The critic network is updated by minimizing the mean squared error of the expected returns from the target critic network and the predicted action-value by the critic network. By minimizing loss, the action-value is as close as possible to the target. The following equations can be used to calculate the target Q value (y) and the critic loss function (J_Q) [47].

$$y = r + \gamma(1 - d) Q'(s', \mu'(s'|\theta^{\mu'})|\theta^{Q'}) \tag{2.14}$$

$$J_Q = \frac{1}{|B|} \sum_B (y - Q(s, \mu(s|\theta^\mu)|\theta^Q))^2 \tag{2.15}$$

Target update: Similar to the actor and critic networks, the target networks are also updated. A "soft update" is performed, in which only a minor fraction of the main networks weights are transferred. It is important to note that target network parameters are not trained, but occasionally synchronized with the parameters of their respective actor and critic networks, as given below:

$$\begin{aligned} \theta^{Q'} &= \tau\theta^{Q'} + (1-\tau)\theta^Q \\ \theta^{\mu'} &= \tau\theta^{\mu'} + (1-\tau)\theta^\mu \end{aligned} \tag{2.16}$$

2.2.4 Multi-Agent Deep Deterministic Policy Gradient

While DDPG is useful in solving problems within complex environments, its usage is confined to single-agent problems. In the real world, the actions of agents could have an impact not only on the environment, but also on other agents. Thus, a variation of DDPG is required to address the multi-agent problem. MADDPG is an actor-critic algorithm that deals with continuous state and action spaces in a problem with multiple agents [35]. The MADDPG operation process of N agents is shown in Figure 2.10. MADDPG proposes centralizing the training operation, where the critic networks use all available information to embed into the actors the dynamics of the environment and rest of the agents. Additionally, from the figure, it is noted that during the operational/testing phase only the actor networks are used based on local information, pertaining to the each respective agent. This is known as the centralized training-decentralized testing approach.

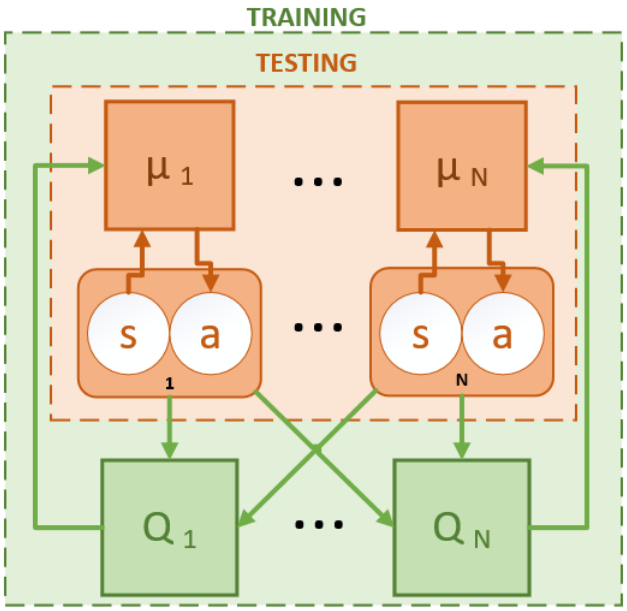


Figure 2.10: Overview of MADDPG Operation

The update processes for the critic, actor, target actor and target critic networks are consistent with that of the DDPG algorithm as explained in Section 2.2.3 with minor modifications to accommodate for the centralized training-decentralized testing approach.

MADDPG, like the DDPG algorithm, utilizes a replay buffer to store the experiences of the agents. Unlike DDPG however, the replay buffer is larger as a result of the multiple agents interacting within the environment. The experiences are once again stored in the form of a tuple (S, A, R, S'), where each of the sets S, A, R and S', represent the state, action, reward and next state values, respectively, of agents i where $i = \{1,...,N\}$. The actor and critic updates in the MADDPG algorithm are given as follows:

Actor Update:

$$\nabla_{\theta^{\mu_i}} J(\theta) = \frac{1}{|B|} \sum_{B} \nabla_{\mu_i} Q_i(S, \mu_i(s_i|\theta^{\mu_i})|\theta^{Q_i}) \nabla_{\theta^{\mu_i}} \mu_i(s_i|\theta^{\mu_i}) \tag{2.17}$$

Critic Update:

$$y = r_i + \gamma(1-d) Q_i'(S', \mu_i'(s_i'|\theta^{\mu_i'})|\theta^{Q_i'}) \tag{2.18}$$

$$J_{Q_i} = \frac{1}{|B|} \sum_{B} (y - Q_i(S, A)|\theta^{Q_i}))^2 \tag{2.19}$$

where $\mu_i = \{\mu_1, ..., \mu_N\}$, $\theta^{\mu_i} = \{\theta^{\mu_1}, ..., \theta^{\mu_N}\}$, $\mu'_i = \{\mu'_1, ..., \mu'_N\}$ and $\theta^{\mu'_i} = \{\theta^{\mu'_1}, ..., \theta^{\mu'_N}\}$ refers to the set of actor networks, actor network parameters, target actor networks and target actor network parameters for all agents, respectively. The *centralized* action-value function, takes as input all actions of all agents, A, state information, S, and outputs the Q-value for agent i.

2.3 Summary

In this chapter, the main concepts required to design the RL agents were presented. First, the operational characteristics of EWHs was reviewed, which included the generation of hot water draw profiles and their electrical load profiles. This was extended to the simulation of profiles for a fleet of EWHs of various sizes. Next, the fundamentals of RL was introduced, with specific concepts such as DRL and MARL explained in brief. Finally, the MADDPG algorithm, which will be used for the research presented in this thesis, was explained in detail.

Chapter 3

Model Architecture of RL Agents and MADDPG Implementation for EWH Energy Management

This chapter presents the agent models for the Residential Aggregator Agent (RAA) and Utility Agent (UA). First, a novel control algorithm using a binning process is discussed. Then, the state, action and reward signals for the RAA and UA are reviewed, along with the agent interaction process under the MADDPG operation. Tables 3.1 and 3.2 list the relevant parameters and variables, respectively, used in this chapter.

Table 3.1: Parameters

Nomenclature	Description
α_μ, $\alpha_{\mu'}$	Learning rate for μ and μ'
β_Q, $\beta_{Q'}$	Learning rate for Q and Q'
C_h^{RES}	Residential consumer hourly electricity price (¢/kWh)
$C_h^{RES(M)}$	(Modified) residential consumer hourly electricity price (¢/kWh)
N^{EWH}	Number of EWHs controlled by RAA
$P^{Pk,forecast}$	Daily forecasted peak load (kW)
$P_h^{forecast}$	Forecasted load at hour h (kW)
$P^{Pk,Actual}$	Daily actual peak load (kW)
ρ_h	HEP or Hourly Energy Price (¢/kWh)
T	Episode duration (hours)

Table 3.2: Variables

Nomenclature	Description
a_h^{RAA}	RAA action signal
s_h^{RAA}	RAA state signal
r_h^{RAA}	RAA reward signal
a_h^{UA}	UA action signal
s_h^{UA}	UA state signal
r_h^{UA}	UA reward Signal
$C_h^{RES(M)}$	Consumer hourly electricity price - UA Modified (¢/kWh)
$P_h^{D,U}$	Utility power demand at hour h (kW)
$P_h^{D,U(x)}$	Utility power demand at hour h w/o EWH load (kW)
$P_h^{D,EWH}$	Modified EWH load from RAA (kW)
$VB_{h,t}^{TES\_actual}$	Actual VB TES at hour h and min t (kWh)
$VB_h^{flex\_ch}$	VB charging limit at hour h (kWh)
$VB_h^{flex\_dch}$	VB discharging limit at hour h (kWh)
VB_h^{flex}	VB flexibility at hour h (kWh)
$VB_{h,t=t'}^{TES\_desired}$	Desired VB TES at hour h and $t' = 60$ (kWh)
$VB_h^{TES\_diff}$	TES difference at hour h (kWh)
VB^{TES_min}	Min allowed TES of VB (kWh)
VB^{TES_max}	Max allowed TES of VB (kWh)

3.1 Overview

The electricity distribution network comprises a multitude of consumers from various sectors such as residential, commercial and industrial, who will likely have varying electricity usage requirements. For instance, the load profile of a residential consumer varies significantly from that of an industrial facility, both in magnitude and pattern. Similarly, the objectives of these consumers can also differ. In addition to traditional loads, the modern distribution grid is becoming populated with consumer smart devices (*e.g.*, smart loads, EVs) and distributed generation (DG) resources (*e.g.*, solar PV, wind and small hydro). These have led to increased complexities in the distribution system by introducing more variability. As the system complexity rises, decentralized and distributed control approaches are being explored as possible ways to manage grid services [50].

AI, and in particular RL algorithms, has proven to be useful in finding the optimal

control policies for a given agent [51]. However, learning control policies in environments with multiple agents is a different, more involved problem. Multiple agents can be used to solve a single task or individual agents may be required to interact in an environment with other agents. Determining the optimal control policies in such scenarios is challenging since the agents can change their behaviour based on the actions of the neighbouring agents. This multi-agent problem can be seen as a game, with the agents being the players of this game. Game theory is often used in problems aiming to select optimal actions for multiple players in a multi-player environment [51]. In the context of the electricity distribution network, the players can be the different components connected to the grid such as, a fleet of EVs, residential houses, DGs, flexible loads or industrial facilities. The manner in which the players are modelled and interact with each other is dependent on their assigned task.

In this thesis, the main goal of the multi-agent game is EWH energy management. A multi-player environment representing the residential loads connected to a distribution feeder will be modelled. This study is a proof of concept and can be extended to larger systems with additional players, representing various power system components. The players will be driven primarily by self interest, with objectives including (but not limited to) energy cost minimization and occupant comfort maximization. A DRL algorithm will be used to model the behaviour of the players, henceforth referred to as "agents".

The multiple agents used in this problem are shown in Figure 3.1. Each agent has its own specific objective, and thus it takes the best actions (based on the learned optimal policy) to maximize its respective reward. The environment is explained in detail in the next section and the agents are defined as follows:

- RAA. The primary task of the RAA is to modify the electricity consumption of a fleet of EWHs based on price signals. In this study, EWHs are considered to be flexible loads, capable of varying their power consumption based on price signals, while ensuring that the quality of service is not compromised. Such a control process requires accurate communication of individual water heater characteristics to the RAA. The thermal energy stored (TES) obtained from the tank water temperature of each EWH tank is aggregated to form a virtual battery (VB). The RAA sends Charge/Discharge signals to the VB, thereby controlling the operation of EWHs.

  For example, if the RAA wants to increase the SoC of the VB, a "charge" signal would be sent to a set of EWHs to turn ON their heating elements. This will result in a rise in tank water temperature and TES, and likely increase the VB SoC (which is not guaranteed as the SoC depends on the aggregate TES of all EWHs in the fleet). The RAA acts in the interest of consumers and aims to reduce costs associated with EWH energy consumption. The RAA is further explained in Section 3.3.

- UA. The UA sends price incentive/penalty signals to the RAA to further influence the behaviour of the EWHs. If the UA seeks to reduce the load at a certain hour, it would send a price penalty signal which would discourage some of the currently charging EWHs to continue their charging, thereby potentially reducing the overall load. The UA acts in the interest of the utility and aims to reduce costs from purchasing power and maximize benefits associated with deferral of infrastructure investments. The UA is further explained in Section 3.4.

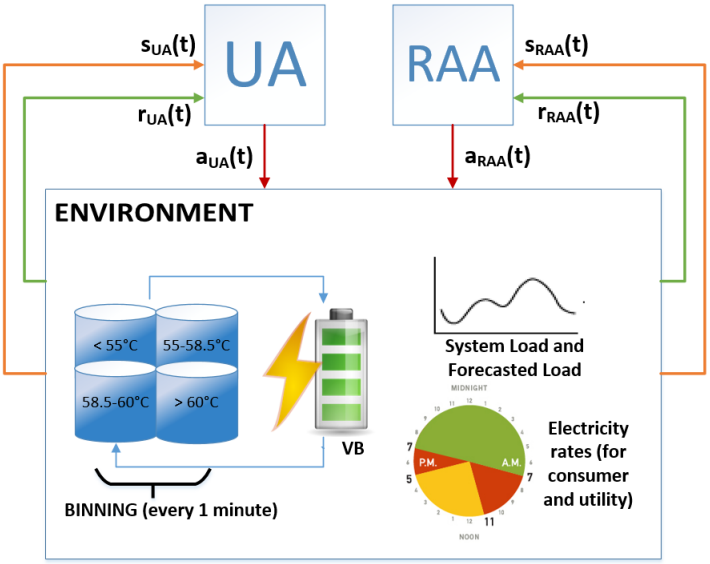


Figure 3.1: Overview: Multi Agent Interaction

3.2 Environment

The environment is a key component for any RL problem in which the agent(s) live and interact. The environment is governed by rules which are referred to as the *environment dynamics*. In the context of this multi-agent problem, the environment dynamics includes load and price datasets and EWH operation. The datasets are required to emulate the system behaviour and to evaluate the benefits (or drawbacks) of the proposed control algorithm. The EWH operation characteristics are discussed in Chapter 2.

Different regions possess different electricity price structures and load profiles. To ensure proper agent training, and thereby accurate execution of the algorithm in the real

world, valid, clean and region-specific data is necessary. Residential load data from a given jurisdiction (in this work, from Summerside, Prince Edward Island, Canada) will be used as the base load profile, while the RAA, through control of EWHs, will generate a modified load profile. Price data includes the electricity tariff rates for consumers and the wholesale electricity market price for the utility. Consumer electricity tariff rates are required to calculate their potential monetary savings while electricity market prices are required to calculate the utility's cost of purchasing electrical energy. As the number of houses increases (in the case of larger feeders, for instance), the potential number of controllable EWHs also increases.

Sending an excess of state signals would make it challenging for an agent to converge to an optimal solution, while on the other hand, a paucity of state signals would leave the agent unaware of the conditions in the environment. A control algorithm via a binning process is employed to address this issue of handling increasing numbers of EWHs in the problem considered in this thesis, and is discussed in Section 3.3.2.

3.3 Residential Aggregator Agent

This section outlines the modelling process and relevant information behind the design of the RAA. Section 3.3.1 explains the operation of the VB from the aggregation of a fleet of EWHs and the resulting load flexibility potential. Section 3.3.2 describes the aforementioned control algorithm employed by the RAA to send control signals to specific sets of EWHs. Lastly, sections 3.3.3, 3.3.4 and 3.3.5 present the state, action and reward signals for the RAA, respectively.

3.3.1 Virtual Battery Operations

As stated in Chapters 1 and 2, the aggregated behaviour and control of EWHs is crucial for providing services to the grid. To that effect, it was also established that simulating the operation of 100 EWHs provides a sufficiently accurate representation of the load demand profile of larger fleets of EWH populations. The RAA will be able to leverage this scalable property during RL training in order to control large populations of EWHs during execution.

When seeking to modify the control of residential appliances such as EWHs, consumer satisfaction is a key priority. In the case of EWHs, satisfaction refers to the availability

of hot water when required. In order to minimize consumer energy cost, electricity usage during certain hours of the day can be modified since time-varying pricing provides opportunities for load shifting or reduction.

A population of EHWs is simulated using the parameters listed in Tables 2.1, 2.2 and the methodology outlined in Chapter 2. The control exerted by the RAA on the EWH population cannot compromise on user comfort (hot water availability). This is accomplished through granular-time (minute-based) control of the EWH heating elements.

The amount of TES in the VB can be calculated by summing the TES in each EWH within the fleet, given as follows.

$$VB_{h,t}^{TES\_actual} = \sum_{n=0}^{N^{EWH}} \frac{mc(\theta_n(t) - \theta_{in})}{3600} \tag{3.1}$$

Thermal energy is energy stored in the form of heat, and is directly proportional to substance mass, specific heat capacity and temperature difference. In this case, the mass and specific heat capacity are those of water, and the temperature difference is between the uniform tank water temperature and inlet water temperature. The values used in this study for these parameters can be found in Tables 2.1 and 2.2. In addition, Table 3.3 presents select EWH temperature values and their corresponding TES (obtained using (3.1)) and SoC, which will be referenced throughout this study.

Table 3.3: TES and SoC of Individual EWH at Select Temperatures

Temperature (°C)	TES - Summer (kWh)	TES - Winter (kWh)	SoC(%)
55.0	10.0	16.8	84.6
57.5	10.8	17.6	91.0
60.0	11.6	18.4	97.3
61.0	11.9	18.7	100

Based on the amount of TES within the VB at a given time ($VB_{h,t}^{TES\_actual}$), it is possible to calculate the charge/discharge limit of the VB, which are dependent on the maximum and minimum TES in the VB, given by the following equations:

$$VB^{TES\_max} = \sum_{n=0}^{N^{EWH}} \frac{mc(\theta_{wh}^{max} - \theta_{in})}{3600} \tag{3.2}$$

$$VB^{TES\_min} = \sum_{n=0}^{N^{EWH}} \frac{mc(\theta_{wh}^{min} - \theta_{in})}{3600} \tag{3.3}$$

The values of θ_{wh}^{min} and θ_{wh}^{max} are 55 °C and 60 °C, respectively, as also given in Tables 2.1 and 2.2. The TES limits ensure that each EWH operates within the acceptable temperature boundaries and that consumers' hot water requirements are not compromised. Accordingly, the *usable thermal energy* is the difference between the maximum and minimum TES in the VB.

In order for the RAA to make an intelligent decision on charging/discharging the VB, their limits, which are calculated on an hourly basis, need to be included in the state signal. For instance, if the VB is in charging mode and approaches 100% SoC at the end of an hour, the discharge limit for the next hour will be large, implying potential to deactivate a significant number of EWHs' heating elements. Similarly, a low SoC of the VB at the end of an hour implies a potential for a large charging limit to increase energy consumption of the EWHs. The RAA will take these factors into account to arrive at the optimal decision. The charging and discharging limits of the VB are obtained using the following:

VB charging limit (in kWh):

$$VB_{h+1}^{flex\_ch} = VB^{TES\_max} - VB_{h,t=t'}^{TES\_actual} \tag{3.4}$$

VB discharging limit (in kWh):

$$VB_{h+1}^{flex\_dch} = VB_{h,t=t'}^{TES\_actual} - VB^{TES\_min} \tag{3.5}$$

where $t' = 60$ min in (3.4) and (3.5) denotes the last minute of hour h. This is because to evaluate the charging and discharging potential of the VB for the next hour, the actual VB TES at the end of the previous hour is required.

The charging/discharging limits of the VB denote the maximum *flexibility*, which can be achieved through the modification of electrical load profile via control schemes while minimizing the impact on consumers and normal EWH operations [52]. The desired VB flexibility at an hour is determined by the RAA action signal. The RAA, during the training process, learns the optimal control policy and sends the appropriate action signal a_h^{RAA}. The desired flexibility is defined as follows:

$$VB_h^{flex} = \begin{cases} a_h^{RAA} VB_h^{flex\_ch}, & \text{if } a > 0 \\ a_h^{RAA} VB_h^{flex\_dch}, & \text{if } a < 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.6}$$

where a_h^{RAA} is a continuous variable in the range [-1, 1]. An action signal $a_h^{RAA} = -1$ or 1 implies a step function (*i.e.* a desire to fully charge/discharge the VB each hour within its usable thermal energy limits). However, a full charge/discharge at certain hours may not be the most intelligent decision by the RAA as it may render certain EWHs at an undesirable state to respond to the consumer's needs. Also, the VB TES, governed by the variable dynamics of EWH operations, may not be able to respond to a full charge/discharge signal, thereby further implying an unintelligent/ineffective RAA signal. Note that in (3.6), the polarity of VB_h^{flex} is identical to that of a_h^{RAA}. Hence a positive a_h^{RAA} is analogous to a charging signal and a negative a_h^{RAA} is a discharging signal.

Given the desired VB charge/discharge flexibility, obtained from (3.6), the desired VB TES (in kWh) at the end of hour h can be calculated as follows:

$$VB_{h,t=t'}^{TES\_desired} = VB_{h,t=1}^{TES\_actual} + VB_h^{flex} \tag{3.7}$$

where $t' = 60$ minutes. $VB_{h,t=t'}^{TES\_desired}$ represents the TES in the VB that the RAA desires to achieve based on the a_h^{RAA} signal; depending on its polarity, $VB_{h,t=t'}^{TES\_desired}$ can be greater or less than $VB_{h,t=1}^{TES\_actual}$, representing a net increase or decrease in the SoC of the VB, respectively.

3.3.2 RAA Control Algorithm Using a Binning Process

All EWHs controlled by the RAA are placed in certain bins based on their tank water temperature. A simple binning process is applied to EWHs such that the RAA can send control signals to the EWHs in specific bins without saturating its state signal. The control algorithm proposed in this work provides the RAA with necessary information which gives an accurate quantitative representation of the environment. The EWH and bin dispersion based on water temperature is given in Table 3.4.

Table 3.4: Classification of Bins Based on EWH Temperature Dispersion

Bin	Temperature (°C)
1	< 55
2	55 - 57.5 (inclusive)
3	57.5 - 60
4	> 60

Utilizing the binning approach, the RAA does not need to know the specifics of each EWH, but rather only their dispersion across the four bins. For instance, out of a fleet of

200 EWHs, if 150 are in bins 3 and 4, this can be a good indicator for the RAA that the VB has a high SoC. Whereas, if the majority of EWHs are in bins 1 and 2, the opposite is true. RAA action signals target only the EWHs within certain bins while the remaining bins operate under default operations. For example, a RAA charge signal would activate the heating elements of EWHs in bins 1 and 2 and a discharge signal would turn off the heating elements of the EWHs in bins 3 and 4. The RAA will have a good idea of the SoC and distribution of TES across the population simply by knowing the number of EWHs in each of the four bins. This means that instead of including temperatures of all EWHs in the RAA state signal, four specific values, one for each bin, can be used to define the ratio of EWHs in each bin to the rest of the population.

In Figure 3.2, the condition selections refers to the following controls to be applied to EWHs; only one condition can be valid at any given time.

- **Condition 1**: $VB_{h,t}^{TES\_actual} < VB_{h,t=60}^{TES\_desired}$. Selected set of EWH heating elements are activated (turned ON). Remaining EWHs operate normally.
- **Condition 2**: $VB_{h,t}^{TES\_actual} > VB_{h,t=60}^{TES\_desired}$. Selected set of EWH heating elements are deactivated (turned OFF). Remaining EWHs operate normally.
- **Condition 3**: $VB_{h,t}^{TES\_actual} = VB_{h,t=60}^{TES\_desired}$. All EWHs operate normally.

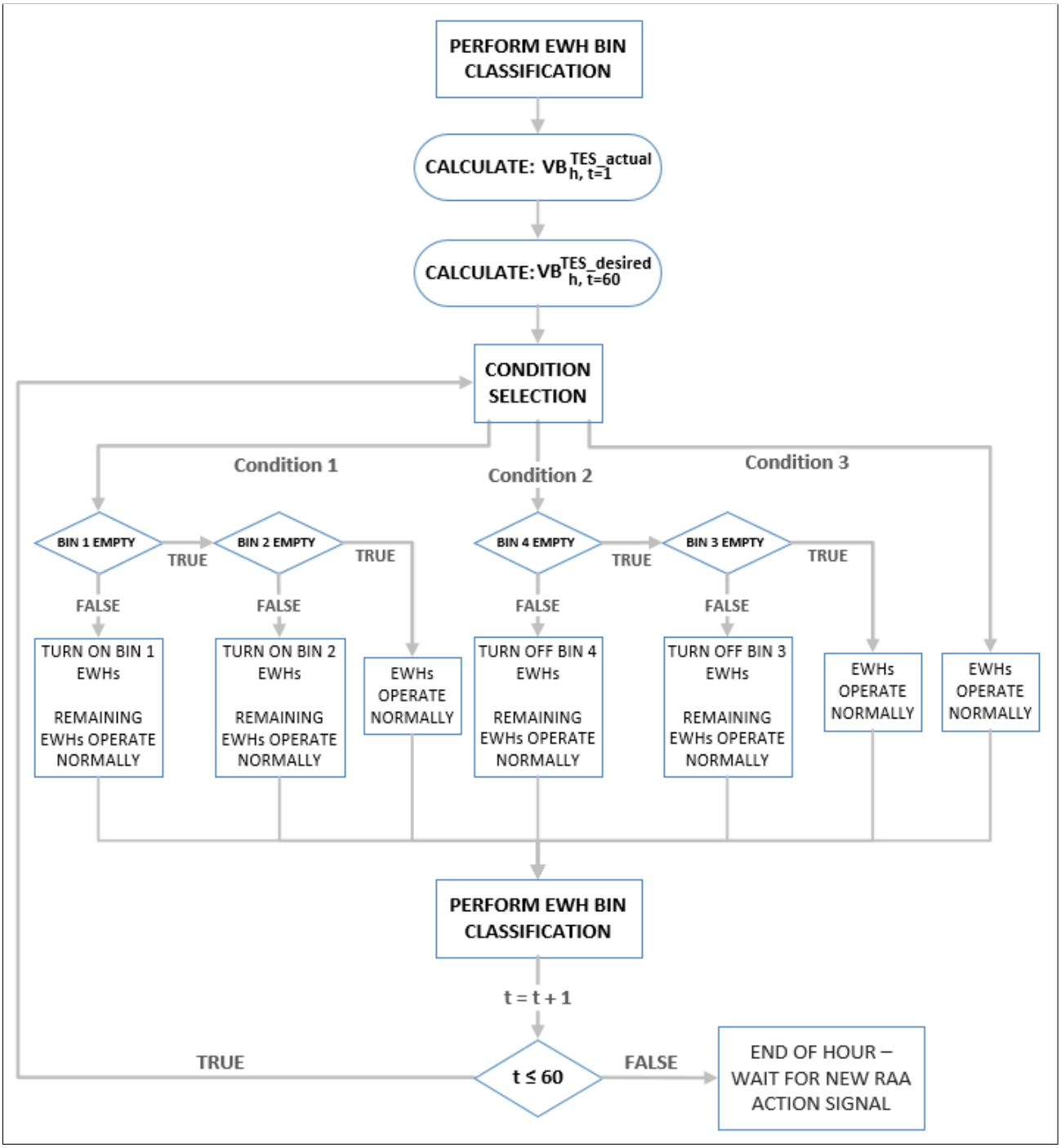


Figure 3.2: RAA Control Strategy Using a Binning Process

3.3.3 State Signal

The state signal components provide the RAA with a representation of the environment and is shown in Figure 3.3. This state signal consists of the following components: time of day, TES distribution (number of EWHs in each of the 4 bins), residential electricity tariff, VB SoC, and VB charge/discharge limit. Parameter tuning of the state signal is necessary to ensure the optimal number of components in the signal.

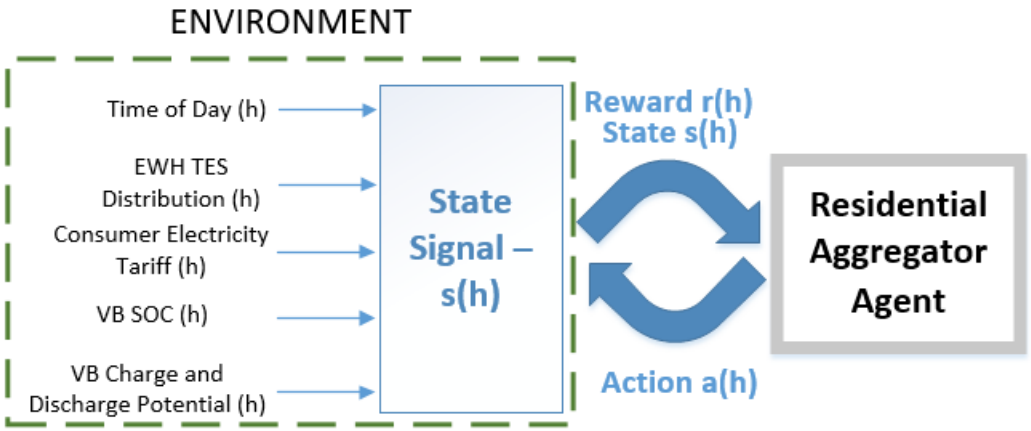


Figure 3.3: RAA State Signal

3.3.4 Action Signal

The RAA sends hourly charge or discharge signals (a_h^{RAA}) to the VB. A charge signal sends a command to the VB to increase the TES and SoC of the battery. The opposite is true for the discharge signal. Over the course of the training process, the RAA learns an optimal behaviour policy which maximizes its rewards. This is done through a trial-and-error process, in which the initial actions are random, until the agent converges to the optimal solution.

3.3.5 Reward Signal

There are three components to the RAA reward signal which are necessary to ensure proper RAA learning: cost of electric energy for charging the EWHs, charge/discharge accuracy and UA incentive. Since consumer cost savings at a given hour is only achieved through a

reduction in electricity consumption, the RAA will be very conservative in sending charging signals to the EWHs. However, reducing electricity consumption during certain hours of the day can be difficult, and more importantly can deter the EWH performance during the following hours. An intelligent agent or controller would anticipate the EWH behaviour and ensure that the tank water temperature is at an acceptable level to provide the necessary hot water, especially during the peak usage times. If the RAA consistently sends discharge signals, even at hours when hot water drawn is very low, the tank water temperature would be towards the lower end of the temperature dead band. As a result, the consumer will be at a higher risk of receiving tepid water, and the power drawn will increase due to the activation of EWH heating elements. Since the price component of the reward function is unable to account for such behaviour, accuracy is deemed a required metric for the RAA reward signal. The electricity cost to the consumers can be defined as follows:

$$C_h^{RAA} = P_h^{D,EWH} C_h^{RES(M)} \tag{3.8}$$

where $P_h^{D,EWH}$ is the aggregated power demand of the EWHs under RAA control at hour h. The accuracy component (ϵ_h) of the reward signal is formulated as follows:

$$\epsilon_h = \begin{cases} -|VB_h^{TES\_diff}| + 10, & \text{if } |VB_h^{TES\_diff}| \leq 10 \\ -2|VB_h^{TES\_diff}| + 20, & \text{if } |VB_h^{TES\_diff}| > 10 \end{cases} \tag{3.9}$$

where $VB_h^{TES\_diff}$ is the difference between the desired and the actual TES in the VB at the end of the hour. In other words, the lower the value of $VB_h^{TES\_diff}$, the more accurate is the RAA.

$$VB_h^{TES\_diff} = VB_{h,t=60}^{TES\_desired} - VB_{h,t=60}^{TES\_actual} \tag{3.10}$$

The function and specific parameters in (3.9) were obtained through initial observation and fine-tuning during the training phases of the RAA and UA. It is noted from (3.9) that when $VB_h^{TES\_diff}$ is below a certain threshold (10 kWh in this study), the accuracy component (ϵ_h) is positive which means the action signal, a_h^{RAA}, is accurate and the RAA is rewarded. The closer $VB_h^{TES\_diff}$ is to zero, the higher is the value of ϵ_h and more accurate is the action signal and higher the reward. When $VB_h^{TES\_diff}$ exceeds 10 kWh, a_h^{RAA} is deemed inaccurate, ϵ_h becomes negative and a negative reward (or a penalty) is imposed. A limit of 10 kWh was deemed sufficient for a VB containing the TES of 100 EWHs. Since the reward signal is only used in the training phase and not required during the execution phase, the specific parameters may remain unchanged as they were designed to accommodate for the training of 100 EWHs as it provides a sufficient representation for large fleets of EWHs.

The third and final component of the reward signal is a penalty/incentive (ϕ_h^{RAA}) associated with the UA penalty/incentive signal. This component is primarily added to accommodate the constant tariff rate structure, as it was noticed that otherwise, the RAA would converge to sub-optimal policies. This penalty/incentive encourages the RAA to charge when there is a price decrease and discharge when there is price increase.

$$\phi_h^{RAA} = \begin{cases} -2P_h^{D,EWH}, & \text{if } C_h^{RES(M)} > C_h^{RES} \\ 2P_h^{D,EWH}, & \text{if } C_h^{RES(M)} < C_h^{RES} \\ 0, & \text{otherwise} \end{cases} \tag{3.11}$$

The three components must be tuned appropriately to obtain the final reward signal. This tuning process is necessary to ensure that one component is not overly prioritized over another. The complete reward signal is given as follows.

$$r_h^{RAA} = aC_h^{RAA} + b\epsilon_h + c\phi_h^{UA} \tag{3.12}$$

It is noted that the three components in (3.12) have different units. The components representing electricity cost, accuracy of the reward signal and the associated penalty/incentive are measured in cents, kWh and kW, respectively. All components are translated into dollars via parameters a, b and c selected arbitrarily and tuned through trial and error. a = -0.01, b = 0.1\$/kWh, and c = 0.2\$/kW.

3.4 Utility Agent

As stated in its name, the UA represents a typical utility in the power system context. The UA is responsible for distribution of electricity and meeting the power demand of consumers. Since the main flexible loads under study are EWHs, the UA seeks to leverage the behaviour of the EWH fleet by sending price incentive signals to the RAA. Figure 3.4 outlines the process used in this study to create a new system demand profile given the modified EWH consumption.

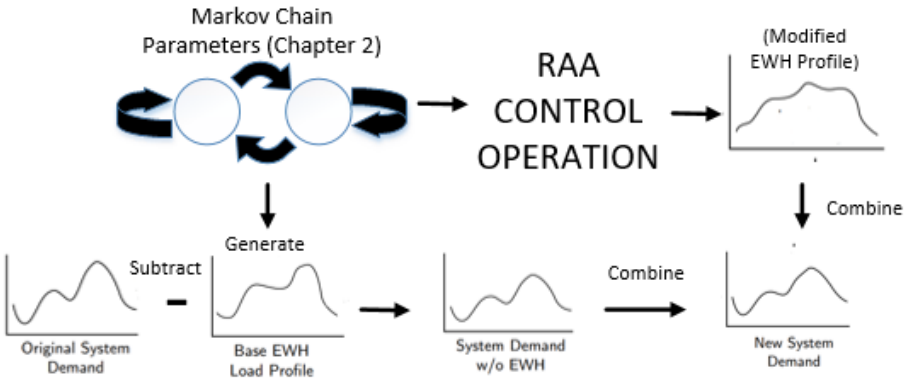


Figure 3.4: EWH Modified System Load Profile

Using the generated Markov chain parameters, a base EWH load profile can be generated. This base load may vary in shape and/or magnitude depending on region, temperature and hot water usage patterns. The RAA uses the same Markov chain parameters but modifies the load profile based on the received charge/discharge signals. Additionally, the UA will attempt to influence the RAA control by sending incentive/penalty signals, thus affecting the resulting load profile. Prior to assessing the impact of RAA and UA modified EWH load profile, the base EWH load profile is subtracted from the original system demand. The system demand (without EWH load) combined with the agent-generated load profile is the resulting system demand profile. The following subsections detail the creation of the state, action and reward signals of the UA.

3.4.1 State Signal

The state signal of the UA comprises: time of day, Hourly Energy Price (HEP), consumer electricity tariff, utility demand, 1, 6, 12, 18 and 24 hour-ahead forecast demand, forecasted daily peak load, and SoC of the VB. Similar to the formulation of the state signal for the RAA, UA state signal parameter tuning is necessary to arrive at the optimal number of signal components. Specific values included in the state signal of the RAA (*i.e.*, EWH bin dispersion and charging/discharging limits) are not necessary in the learning process of the UA (Figure 3.5).

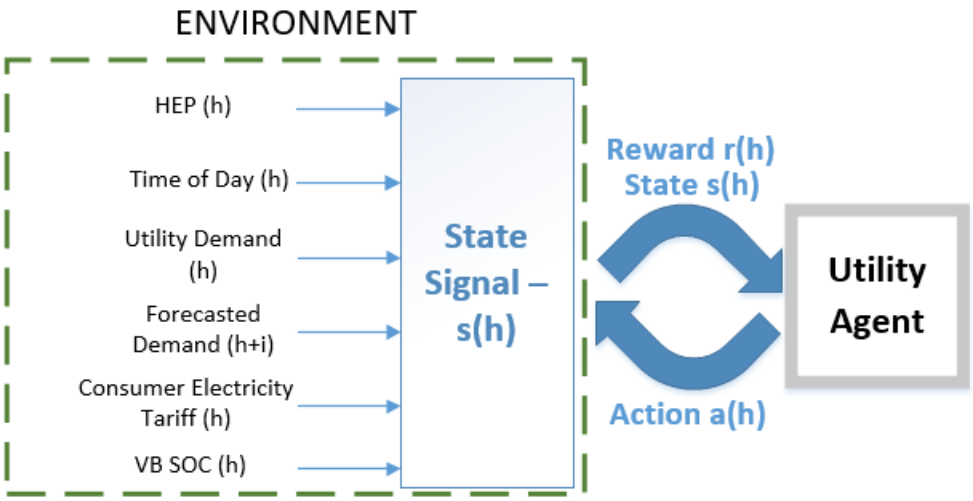


Figure 3.5: UA - State Signal

3.4.2 Action Signal

The UA will send hourly action signals, a_h^{UA}, to the RAA in the range [-1, 1], where a_h^{UA} is a continuous variable. These action signals are scaled by a multiplier, which in this study is considered to be 3 ¢/kWh. This implies, the utility can provide a price incentive ranging from -3 ¢/kWh (reducing C_h^{RES} by 3¢/kWh) to +3 ¢/kWh (increasing C_h^{RES} by 3 ¢/kWh). The choice of the scaling factor of 3 ¢/kWh was based on trial-and-error and previous reported works. Note that in [26] a 5 (euro) ¢/kWh incentive is used while [53] uses an incentive of 3 ¢/kWh; thus the chosen range is reasonably appropriate. The objectives of the UA are minimization of the operation cost and peak load reduction via load shifting, for which different hourly price signals are generated. The RAA, designed to control the EWH usage in a manner which minimizes consumer energy consumption cost, will adjust its behaviour accordingly.

3.4.3 Reward Signal

The reward function of the UA comprises the net revenue of UA and peak demand reduction. Net revenue of the UA is the difference between revenue obtained from EWH energy consumption and cost of purchasing electrical energy, as given below.

$$\Omega_h^{UA} = C_h^{RES(M)} P_h^{D,EWH} - \rho_h P_h^{D,EWH} \tag{3.13}$$

where $C_h^{RES(M)}$ is the modified tariff rate for the residential consumer.

$$C_h^{RES(M)} = C_h^{RES} + 3a_h^{UA} \tag{3.14}$$

The multiplier associated with the second term in 3.14 denotes the 3 ¢/kWh incentive considered in this study, as discussed earlier. In order to minimize the peak demand, the UA will be penalized/rewarded (denoted by variable ψ_h) when the modified system load ($P_h^{D,U}$) is within certain thresholds.

If $P_h^{D,U} \geq 0.91P^{Pk,forecast}$:

$$\psi_h = \begin{cases} a(P_h^{D,U} - P_h^{forecast}), & \text{if } P_h^{D,U} \geq P_h^{forecast} \\ b(P_h^{D,U} - P_h^{forecast}), & \text{otherwise} \end{cases} \tag{3.15}$$

Else:

$$\psi_h = \begin{cases} c(P_h^{D,U} - P_h^{forecast}), & \text{if } P_h^{D,U} \geq P_h^{forecast} \\ d(P_h^{D,U} - P_h^{forecast}), & \text{otherwise} \end{cases} \tag{3.16}$$

where $P_h^{D,U}$ is the agent-modified load profile for a given hour.

$$P_h^{D,U} = P_h^{D,U(x)} + P_h^{D,EWH} \tag{3.17}$$

where a = -0.3, b = -0.1, c = 0.3, d = 0.5. Parameters a, b, c and d from (3.15) and (3.16) all have units of \$/kW, so that all components of the reward signal are translated into dollars. The two components must be tuned appropriately to obtain the final reward signal. This tuning process is necessary to ensure that one component is not overly prioritized over another. The complete reward signal is given as follows:

$$r_h^{UA} = 0.01\Omega_h^{UA} + \psi_h \tag{3.18}$$

The parameters used for the UA reward signal were obtained through trial and error. Since the UA has multiple objectives in minimizing peak demand and net revenue maximization, each reward signal component had to be scaled appropriately so that both objectives can be met.

3.5 Interaction of RAA and UA Through MADDPG Algorithm

The MARL algorithm employed in this study is a variation of the DDPG algorithm, MADDPG. MADDPG utilizes a centralized training and decentralized testing approach. As outlined in Chapter 2, each agent in the DDPG/MADDPG algorithm is modelled with four neural networks (actor, critic, target actor and target critic networks). When dealing with multiple agents, each with its own objective, the agents have to make informed decisions, which are simultaneously affected by the decisions of other agents operating in the environment. During the "centralized training" the critic networks of each agent receive the state and action signals of the other agents. Since the role of the critic network is to evaluate the actions of the respective actor networks, obtaining relevant information from the operations of other agents is necessary to enhance the agents' learning process. This is shown in Figure 3.6 which presents the interaction of the two agents considered in this study, the RAA and UA.

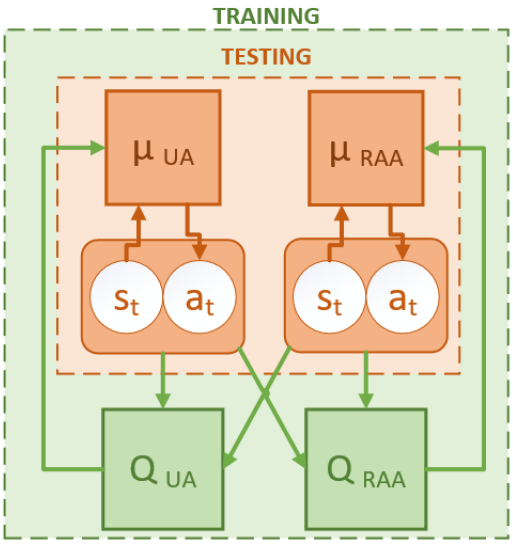
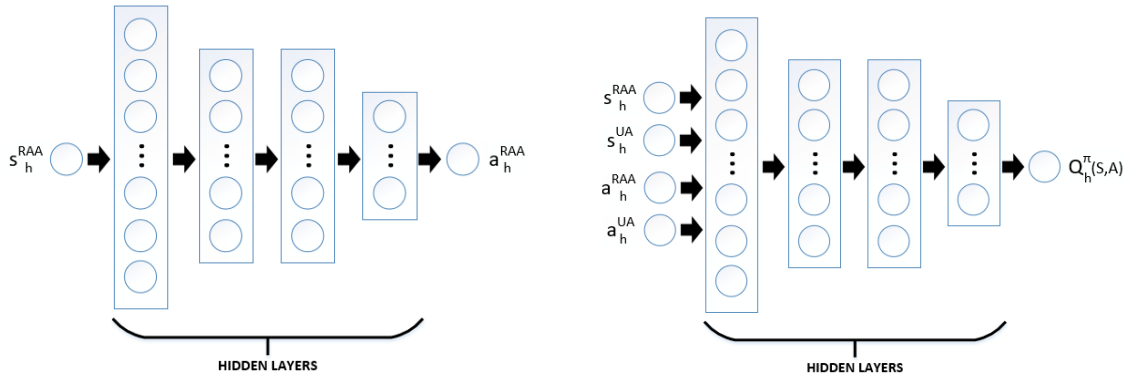


Figure 3.6: Interaction of RAA and UA through MADDPG Algorithm

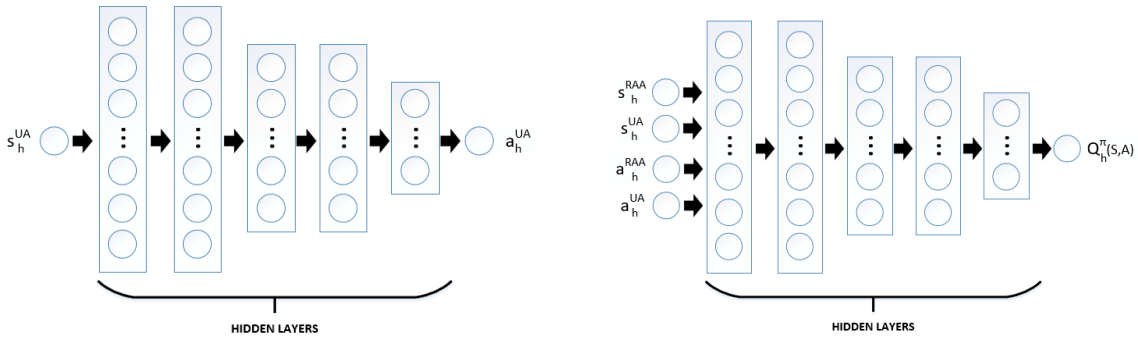
In Figure 3.6, μ_{RAA} and μ_{UA} are the actor networks for the RAA and UA, respectively and Q_{RAA} and Q_{UA} are the critic networks for the RAA and UA, respectively. The target actor and critic networks for both agents (not shown in Fig. 3.6) are represented by μ'_{RAA}, μ'_{UA}, Q'_{RAA}, and Q'_{UA}.

In the context of this study, the RAA and UA critic networks are provided with the state and action signals of the other. During the training process, the agents interact with the environment (and each other) to maximize their respective rewards. As a result of the centralized training methodology, the agents are able to identify patterns of the other, and learn behaviour strategies to respond accordingly. During the testing phase however, the RAA and UA operation is no longer centralized. The RAA will continue receiving hourly price signals from the UA, and based on what it learned during the training process, is expected to react in an intelligent manner. The same is true for the UA, which will receive modified EWH load profiles from the multi-agent operation, and is expected to respond so as to maximize the desired rewards. The actor and critic networks of the UA and RAA are shown in Figures 3.7 and 3.8, respectively.



(a) Actor Network Architecture (b) Critic Network Architecture

Figure 3.7: Actor and Critic Networks for RAA



(a) Actor Network Architecture (b) Critic Network Architecture

Figure 3.8: Actor and Critic Networks for UA

Figures 3.7 and 3.8 presents the architectures of the actor and critic networks and depicts how centralized training and decentralized testing are accomplished in the MADDPG algorithm. In Figures 3.7(b) and 3.8(b), both the critic networks have inputs of state and action signals from the agents and outputs the Q-value for its respective agent's actor network. The actor networks, however, only accepts the state signal from their respective agents as inputs. The hidden layer sizes for the actor and critic networks differ for the two agents. The RAA networks have four hidden layers of 128, 64, 64, 32 neurons while the UA networks have 5 hidden layers of 128, 128, 64, 64, 32 neurons.

3.5.1 Agent Training

Before an RL agent can be confidently used in the real world, it has to be tried and tested. The available datasets used in this study in evaluating the created models is divided into training and testing sets in the following ratio: 80% for training and 20% for testing. In the context of RL, the fine tuning and performance evaluation is accomplished during the training process. The environment, described in Section 3.2, is where the agents in this study (RAA and UA) learn to make the optimal behavioural choices. The following considerations are taken into account for this problem.:

- Different regions may have different electricity tariff rates for consumers. This is especially true across different provinces in Canada, for example, Ontario residential consumers pay a TOU rate while New Brunswick consumers pay a fixed rate. Because of these differences, proper and relevant data is required prior to real-world execution. For instance, when employing the agents in the New Brunswick power grid, price data specific to New Brunswick is used to train the agent. The details and intricacies of the model remain unchanged and transferable across all regions, but this distinction in data is required during the training process.

- Electricity consumption patterns are significantly dependent on seasonal temperature patters. This is especially true in Canada where the winter and summer months play a significant role in electricity usage. Due to these electricity consumption variations, the agents are trained on summer data prior to execution and testing in summer months, and similarly, trained on winter data prior to execution and testing in winter months. The details and intricacies of the model remain unchanged and transferable across all regions, but this distinction in data is required during the training process.

- Due to the trial-and-error nature of RL, the agents need to continuously interact with the environment throughout multiple episodes. An episode is the length of the

simulation at end of which the system ends in a terminal state. In video games, the episode duration could represent the time a video game character is alive. In this EWH energy management problem, the episode duration is set as 336 hours (two weeks), established through trial-and-error. Month-long episode durations would significantly lengthen the training process and day-long episode durations were found to be too short for the agent to learn the optimal control policies.

3.5.2 Agent Testing

Once the agents are trained, the testing/execution is rather straight-forward. Though the models are only trained for 336 hour simulations, the learned behaviours permit the execution of the algorithm for longer durations. Ten randomly selected intervals are chosen to evaluate agent performance. As previously mentioned, as a result of the variations in price and electricity consumption across the various regions and seasons, specifically trained models should be used on a region to region and season to season basis. For instance, a model trained using Ontario summer data should not be used to execute a simulation in New Brunswick as this would lead to inconsistent results.

3.6 Summary

In this chapter, the concepts explained in Chapter 2 were used to create the agent models. First, the problem was defined as a multi-agent task and the interactions between the agents and the agents with the environment was discussed. Second, the RAA model was built incrementally. The proposed RAA control algorithm via a binning process to control a population of EWHs was explained, and the state, action, and reward signals for the RAA were defined. Thereafter, the UA agent model was built incrementally, and its state, action and reward signals were clearly outlined. The objectives and rationale behind constructing these two specific agents were also discussed. The UA and RAA interactions were simulated using the MADDPG algorithm in the EWH energy management problem. The centralized training / decentralized testing approach and the NN architecture for actor, critic, target actor and target critic networks for each agent were presented.

Chapter 4

Application of MADDPG Algorithm for EWH Energy Management of Residential Consumers in Ontario, New Brunswick and Quebec

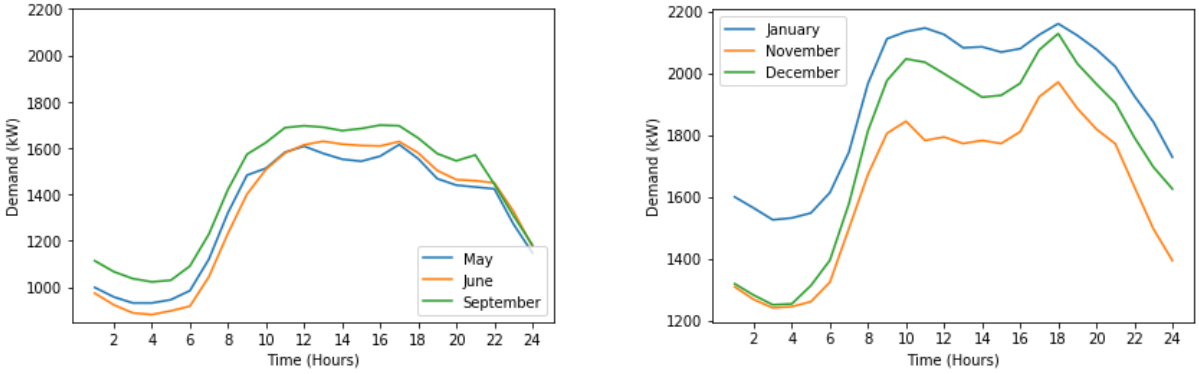
Based on the RAA and UA models and the MADDPG algorithm presented in Chapter 3, this chapter evaluates the performance of the two interacting agents considering real price data from three provinces of Canada and load data from Summerside, Prince Edward Island (PEI), Canada. The residential electricity tariff rates applicable in the provinces of Ontario, New Brunswick and Quebec are examined and utilized for the different scenarios. Based on the test results over multiple scenarios, consumer and utility savings are determined and simulations outlining the modified load profiles of the EWH loads are presented.

4.1 Input Data

Due to the data-driven nature of RL problems, clean, accurate and reliable data is required to adequately train the agents. To simulate the behaviour of an electric utility and obtain results from the agent models designed in Chapter 3, the following datasets are used:

- Base Load Profile: Historical load data of 2016, from Summerside, PEI is used to represent the generic base load profile of a residential feeder. Figure 4.1 presents the

base load profiles of select days in summer and winter months. It is noted that the occurrence and magnitude of peaks vary with the season, and this is accounted for while training the RL agent, as mentioned in Sections 3.5.1, 3.5.2. In addition to real-time data, the UA also requires load forecasts. Being aware of load demand over the next hours is a reasonable expectation from an UA responsible for purchasing power from generating sources and supplying to consumers. Accordingly, 10-15% noise is added to the base load data to generate load forecast values.



(a) Summer (b) Winter

Figure 4.1: Base Load Profiles

- The installation cost of the EWH controller (C^{EWH}) is assumed to be \$100. Examples of EWH controllers with prices ranging from \$70-\$200 are described in [54]. In the coming years, when EWHs are more commonly used as flexible loads and controllers are installed at the manufacturing stage, their prices are expected to drop significantly. The lifetime of the EWH controller is assumed to be five years.

- The Hourly Ontario Energy Price (HOEP) of Ontario, from 2016 [55] and the Final Hourly Marginal Cost (FHMC) in New Brunswick [56] from 2012 are used to calculate the electricity costs to the respective utilities of these provinces. The year 2016 had 366 days and due to unavailability of FHMC data in 2016, the price data from 2012 is used.

- TOU rates in Ontario [57] and constant tariff rates in Quebec [58] and New Brunswick [59] are required to calculate the electricity cost of consumers in their respective regions. The electricity TOU periods considered in this study, for Ontario, are taken from IESO website [57] and shown in Figure 4.2.

Figure 4.2: Ontario Electricity TOU Periods

In Ontario, during low demand periods, electricity is primarily supplied by cheap, base-load sources such as nuclear and large hydroelectric stations. As demand rises in the morning when people turn on their household appliances and businesses commence operations, additional and typically more expensive sources of power, like natural gas-fired generation plants, are required to meet the increased demand, and such chronological variation of demand takes place over the day. The TOU tariff rate, designed for Ontario, reflects this variation of the cost of generation over a daily load cycle, and which can also vary seasonally, as evident from the summer and winter TOU rates.

Quebec is one of the largest electricity producers in Canada, with the majority of its power coming from the province's abundant hydroelectricity sources (over 130,000 rivers and streams). This rich supply of hydroelectric sources contributes to Quebec having the lowest electricity rates in North America [60].

New Brunswick generates its electricity from a combination of sources. In 2016, nuclear was the primary source of electricity, followed by hydro and coal at 29.9%, 20.6% and 20.7% of New Brunswick's total generation, respectively [61].

The winter and summer electricity rates used in this study for the three aforementioned provinces of Canada are provided in Tables 4.1 and 4.2.

Table 4.1: Winter Electricity Rates (November 1 - April 30)

Region	Off-Peak ($/kWh)	Mid-Peak ($/kWh)	Peak ($/kWh)
Ontario (Weekdays)[1]	0.105	0.15	0.217
Quebec[2]	0.095		
New Brunswick	0.1138		

Table 4.2: Summer Electricity Rates (May 1 - October 31)

Region	Off-Peak ($/kWh)	Mid-Peak ($/kWh)	Peak ($/kWh)
Ontario (Weekdays)[1]	0.105	0.15	0.217
Quebec[2]	0.071		
New Brunswick	0.1138		

The next section outlines the test scenarios presented in this study which simulate the execution of the RL agents in various regions over the year, thereby evaluating agents' performance considering factors like price and load variations.

4.2 Test Scenarios

Based on the feeder load data used in this study, the maximum hourly load is about 2.2 MW. The number of EWHs in a residential feeder can be calculated from the number of houses supplied by the feeder and the percentage of those equipped with EWHs. A diversity factor of 0.5 [62], peak demand of houses from different Canadian regions [63] and province-wide EWH penetration [15] are used to calculate the number of houses within a residential feeder. Appendix A outlines the method used in calculating the number of houses and EWHs within a residential feeder in Ontario; the same process has been extended to New Brunswick and Quebec. Table 4.3 provides a summary of household peak demand, number of houses and EWHs connected, as considered in the case studies for the three Canadian provinces.

[1] In Ontario, off-peak rates apply for weekends and statutory holidays.

[2] Quebec utilizes a two-tiered tariff for its residential consumers. Consumers are charged first-tier price for energy consumption up to a certain amount, and second-tier price for any additional consumed energy. Since energy consumption is noticeably higher in winter months, the second-tier price is applied (as a constant rate) during winter months and first-tier price is applied (as a constant rate) during summer months.

Table 4.3: Case Study Details on EWHs Connected to a Feeder

Region	**Household Maximum Demand (kW)**	**# of Houses**	**% of EWHs**	**# of EWHs**
Ontario	7.2	611	21, 30, 50	128, 183, 305
New Brunswick	9.2	478	92	439
Quebec	15	293	93	272

At approximately 21% of EWH penetration in the residential sector, the province of Ontario has a rather low percentage share of EWHs as compared to the other two provinces. The majority of households in Ontario (almost 75%) utilize natural gas to fuel their domestic water heaters. This is due to the increased presence of gas lines in urban areas and a relatively high electricity cost. The increased reliance on natural gas also reduces household electricity peak in Ontario, relative to New Brunswick and Quebec. In rural areas however, natural gas lines are less common, thereby making electricity the more preferred fuel choice for domestic water heaters. Therefore the simulations for Ontario assumes a rural region. Different penetrations of EWHs (21%, 30% and 50%) are considered to evaluate the performance of the RL agents on the Ontario grid. Since majority of households in Quebec and New Brunswick already have high EWH penetration levels (93% and 92% respectively), no additional penetration scenarios are considered.

4.2.1 Utility Capital Deferment

Increased EWH penetration will result in increased electricity demand seen by the transformer, which will incur additional capital costs. Appropriate EWH energy management schemes can defer such additional capacity requirements by the utility. Utility capital deferment refers to the postponement of power system capital expenditure to be made by the utility, which can be accomplished through peak reduction via DSM techniques. The following approach of calculating expected deferral of investments can be found in [53].

The Peak Reduction Ratio (PRR) is the ratio of system peak demand (with the modified EWH load) to the original/unmodified system load. This is calculated as follows:

$$PRR = \frac{P^{D,EWH}}{P^{D,U}} \tag{4.1}$$

Using the PRR, the expected deferral of investment ($E[\Delta N]$), in years, is calculated as follows:

$$E[\Delta N] = \frac{\log \frac{1}{PRR}}{\log(1+g)} \tag{4.2}$$

where g is the annual growth rate of the system load, assumed 1.5% in this work. Lastly, the benefit accrued to the utility from the investment deferral is calculated as follows:

$$E[B] = C^{Inv}(1 - (\frac{1+i}{1+d})^{E[\Delta N]}) \tag{4.3}$$

where C^{Inv} represents the capital cost of the distribution side equipment, i is the inflation rate (assumed 1.1%) and d is the discount rate (assumed 10%). With a peak load of 2.2 MW for the residential feeder, used in this study, equipment capacity of 2.5 MW (C_{pk}^{Inv}) can be considered appropriate. For a base investment cost C_b^{Inv} of 100 \$/kW [64], the total capital investment cost C^{Inv} is calculated as follows:

$$\begin{aligned} C^{Inv} &= C_b^{Inv} \cdot C_{pk}^{Inv} \\ &= \frac{\$100}{kW} \cdot 2500\, kW \\ &= \$250,000 \end{aligned} \tag{4.4}$$

4.2.2 Comfort Index

As previously mentioned, the control of flexible loads should not compromise on consumer comfort/preference. In the case of EWHs, comfort corresponds to the availability of hot water during times of hot water draw. User comfort is defined as the percentage of time that the water outlet temperature is ≥ 50 °C for simulation period (1 day or 1,440 minutes) [39]:

$$Comfort = \frac{mins\ where\ outlet\ temperature \geq 50\ ^\circ C}{1,440} \tag{4.5}$$

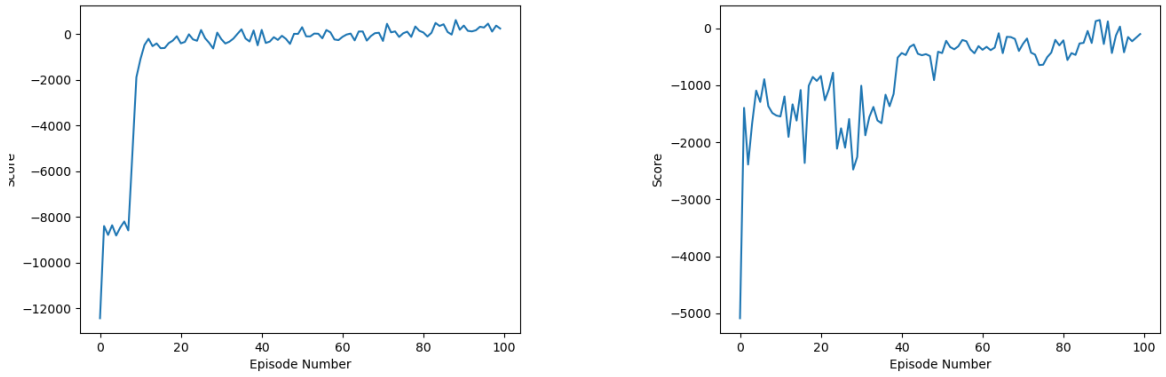
4.3 Results

The training and testing of the simulations is carried out on a Windows 10 desktop computer with Intel (R) Core (TM) i7 - 4770 CPU 3.40 GHz and 16.0 GB RAM. Python

3.7.6 (64-bit) and TensorFlow 2 is utilized to create the necessary models. The rest of this section presents the optimal results obtained from executing the MADDPG algorithm considering the three Canadian regions outlined in Table 4.3.

4.3.1 Case Study: Ontario

Convergence of the RL process of the agents is one of the metrics used to evaluate the effectiveness of the model. In practice, an RL algorithm is said to converge when the learning curve becomes flat and ceases to increase. Figure 4.3 illustrates the convergence behaviour of the RAA and UA for the winter test scenario of Ontario.



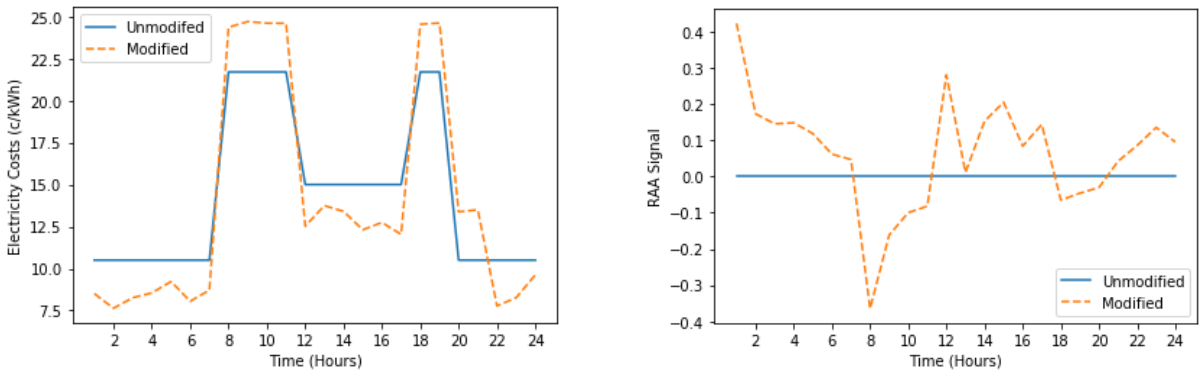
(a) RAA (b) UA

Figure 4.3: Convergence of the Agents' Learning Process

The progressive flattening of the learning curves, after approximately 60 episodes of training, is evident in the diagrams. It is important to note that since the reward signal of both agents comprises multiple objectives, the y-axis (score) does not represent a single metric. For instance, the UA reward signal includes peak reduction and cost minimization components which are scaled appropriately so that the agent is able to consider both objectives. Agent convergence does not always imply the discovery of an optimal behavioural policy, as convergence may have occurred prematurely with a sub-optimal policy. Thus, the performance of the agent is evaluated based on the convergence as well as the results obtained during the testing/execution phase and further model adjustments such as hyper parameter tuning and tuning of state and reward signals is carried out.

As mentioned in Chapter 3, the main objective of the RAA is to control a fleet of EWHs via a binning process, all the while seeking to minimize energy costs for consumers and ensuring that hot water is available when needed. The UA objectives include peak reduction and net revenue maximization which are attained through the leveraging of the modified EWH load profile through incentive/penalty signals sent to the RAA. The results from the best test simulations from the RAA and UA operation are presented in the latter sections.

Residential electricity consumers in Ontario are charged a TOU electricity tariff. The specific rates for off-peak, mid-peak and peak times are given in Tables 4.1 and 4.2. Due to the nature of the TOU tariff structure, electricity consumers can reduce their cost by shifting their energy usage from peak/ mid-peak to off-peak times. To further encourage the shift of electricity usage (for peak reduction and net revenue maximization of the utility), the UA sends an incentive signal to participating EWHs which increases/decreases their hourly electricity tariff. Figure 4.4 depicts the operation of the UA and RAA on a winter day for 21% EWH penetration.



(a) TOU Tariff and the UA Modified Tariff

(b) RAA Action Signal

Figure 4.4: Action Signals of RAA and UA for Winter Simulations

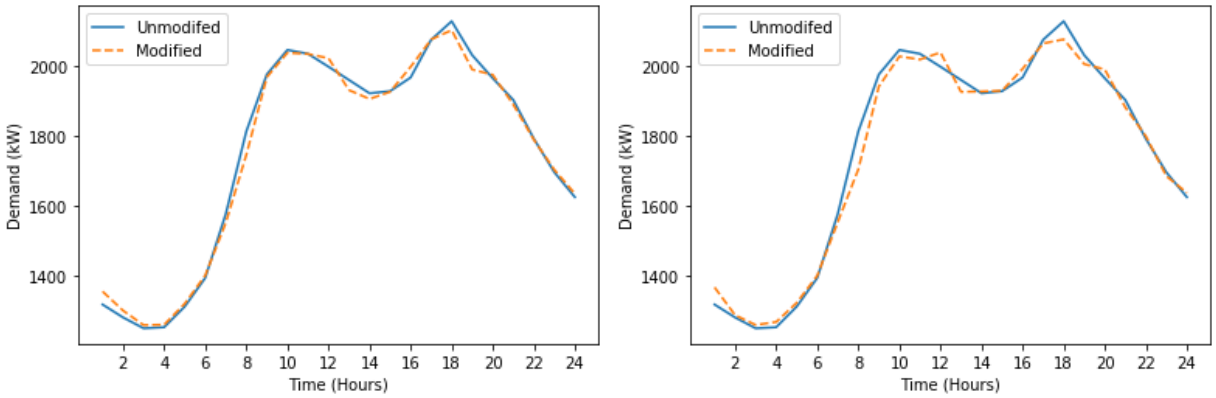
A distinct increase in electricity tariff rates during the morning and evening peak times and a decrease during other times can be observed in Figure 4.4(a). Thus, it can be inferred that the UA, aware of the environment conditions (through the carefully designed state signal), is sending specific incentive signals in an attempt to alter the RAA behaviour, and by association, the behaviour of the fleet of EWHs.

The RAA handles the shifting of EWH energy consumption by activating or deactivating the device heating elements of select EWHs by sending hourly charge or discharge

signals. From the RAA model in Chapter 3, it is noted that RAA action signals are continuous variables in the range [-1, 1], where negative and positive signals imply discharging and charging of the VB, respectively. Figure 4.4(b) illustrates the RAA action signals for the same day of the aforementioned UA price signal.

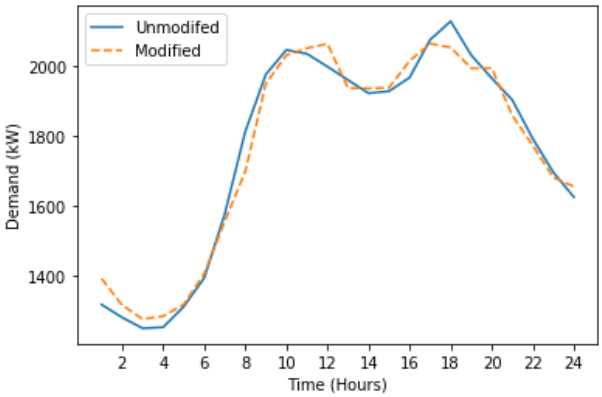
Positive RAA signals, in Figure 4.4(b), in the early hours of the day indicate the RAA's desire to increase the VB TES before the morning peak. The reduced electricity tariff (Figure 4.4(a)) during these hours, further encourage this behaviour. Positive RAA signals in the afternoon hours also indicate the RAA's desire to charge the VB prior to the evening peak. The negative RAA action signals occur primarily during the peak hours in order to discharge the VB. The magnitude of the RAA signals are dependent on the objectives established through the reward signal, namely cost minimization and charge/discharge potential. Lower magnitude signals, such as the discharge signals in the evening peak versus morning peak, is likely because of the lower discharge potential.

The resulting load profiles from multi-agent operation and the impact of the different penetrations of EWHs during the winter simulations can be observed in Figure 4.5.



(a) EWH Penetration of 21.1%

(b) EWH Penetration of 30%



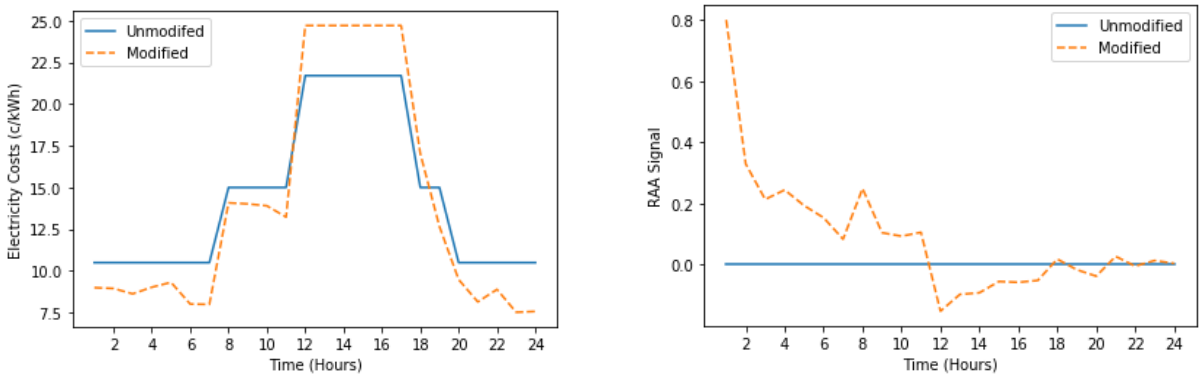
(c) EWH Penetration of 50%

Figure 4.5: Winter Load profile with EWH Operation

As EWH penetration increases, certain patterns can be observed in Figure 4.5- most notably, a gradual reduction in peak load. In the particular day simulated above, the peak load occurs during the evening (with a lesser peak in the morning), and the RAA controls the EWHs in a manner which shifts the peak load to off peak times. Results from the winter simulation indicate a 6% reduction in EWH energy consumption during peak hours; peak hours in this study correspond to the hours associated with the Ontario TOU rates. As a result of the reduced consumption during peak, there is approximately a 3.5% increase in EWH energy consumption during non-peak hours. The remaining shifted load can be attributed to uncertainties in draw profiles and reduced standby heat losses from the modified EWH charging behaviour. Also, note that the simulations presented in Figure 4.5 do not result in significant rebounds during the periods following peaks. The

"rebound effect," or cold load pick up (CLPU), refers to the period in which EWHs under DSM control seeks to restore their thermal energy levels by increasing energy consumption [65]. The rebound, normally associated with EWHs under DR action, is well mitigated due to RAA and UA control.

The rest of the subsection discusses the multi-agent operation in Ontario under summer conditions. Figure 4.6 depicts the operation of the UA and RAA on a summer day for 21% EWH penetration.



(a) TOU Tariff and the UA Modified Tariff

(b) RAA Action Signal

Figure 4.6: Action Signals of RAA and UA for Summer Simulations

Similar to the patterns observed in Figure 4.4, increased electricity tariff rates can be observed during the summer peak times, indicating a push for reduced EWH energy consumption. The RAA operation under summer conditions is illustrated in Figure 4.6(b) and it can be observed that the RAA sends positive/charging action signals in the morning leading up to the afternoon peak. Relatively high positive RAA signals, especially in the low-cost early hours of the day, indicate the RAA's desire to charge the VB and minimize cost. The negative/discharge signals indicate the RAA's desire to discharge the VB during the peak times, with the intention of reducing EWH energy consumption, thereby reducing consumer cost.

The resulting load profiles from multi-agent operation for the 21% EWH penetration scenario on a randomly selected summer day can be observed in Figure 4.7.

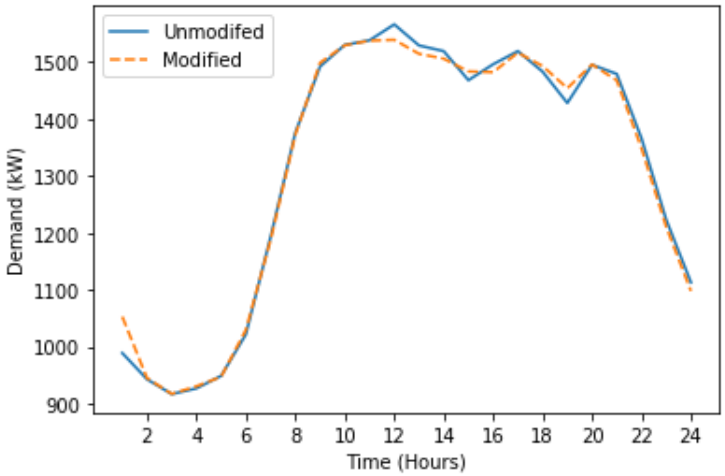


Figure 4.7: Summer Load Profile with EWH Operation

It is evident that the peak reduction potential in summer is noticeably less than in winter as seen in Figures 4.5 and 4.7, the main reason being the non-coincidental nature of EWH electricity demand and the grid peak. EWH load peaks, as shown in Figure 2.6(b), occur at hours 6-8 and 19-21, whereas the grid peak demand in the summer occurs in the afternoon (typically between 11-17 hours as shown in Figure 4.1(a)). During the afternoon, EWH energy consumption is significantly lower than during its peak times, thus rendering EWHs (almost) ineffective in providing peak reduction services. Based on the summer test runs, a 3.9% increase in EWH energy consumption during non-peak hours is noted and 2.3% of the total load is shifted from peak hours to non-peak.

By application of the proposed MADDPG algorithm to the operation of the UA and RAA, the residential consumers can benefit from cost savings accrued to them through optimized EWH control. In this work, the cost savings are assumed divided by the RAA equally amongst all participating EWHs. In practical implementation however, an EWH controller would enable the RAA and UA to be aware of the energy consumption patterns of each participating EWH, thereby accounting for electricity charged/metered individually per house. This would result in cost savings being dependent on the individual contribution of EWHs. Consumer cost savings for different penetrations of EWH, as defined by Cases 1 to 3, are presented in Table 4.4. The winter and summer savings provided in the table are the best savings from the test simulations and the annual savings are calculated by extrapolating the optimal savings with the assumption that winter and summer periods last for six months each.

Table 4.4: Cost Savings to Individual Consumer from EWH Operation

	Winter ($/day)	Summer ($/day)	Annual Savings ($)
Case 1 (21%)	0.174	0.096	49
Case 2 (30%)	0.183	0.097	50
Case 3 (50%)	0.189	0.092	50

The net present value (NPV) of the savings can be calculated from the annual savings, as follows [66]:

$$P = PMT \cdot \frac{1-(1+d)^{-n}}{d} \tag{4.6}$$

where PMT is the dollar amount of each annuity, d is the discount rate (10%) and n is the number of periods in which payments are received. Using (4.6) and annual savings from Table 4.4, the NPV of savings to a consumer, for the three cases, are \$185, \$189 and \$190, respectively. Assuming an EWH controller cost of \$100, the final savings to a consumer in the three cases are \$85, \$89 and \$90, respectively.

The utility is also expected to benefit from the proposed EWH operation – first due to the change in its net revenue from modified EWH operation, net of its cost of purchasing electricity from the grid, and second, from deferment of investment costs through peak shaving. However, as noted from Table 4.4, there are increased savings for consumers from the proposed EWH operations, which imply reduced revenue earnings for the utility. The utility's net revenue for a simulation (of 336 hours) considering controlled EWH operations ($\Omega^{U(M)}$) is expressed as follows:

$$\Omega^{U(M)} = \sum_{h=1}^{336} (C_h^{RES(M)} P_h^{D,EWH} - \rho_h P_h^{D,EWH}) \tag{4.7}$$

where $C_h^{RES(M)}$ is the UA modified electricity tariff at hour h applicable to the EWH consumers while ρ_h is the electricity market price at hour h. On the other hand, the utility's net revenue for a simulation (of 336 hours) considering uncontrolled EWH operation (Ω^{U}) is expressed as follows:

$$\Omega^{U} = \sum_{h=1}^{336} (C_h^{RES} P_h^{D,EWH} - \rho_h P_h^{D,EWH}) \tag{4.8}$$

where C_h^{RES} is the original (unmodified) electricity tariff at hour h. The first terms in (4.7) and (4.8) represents the utility's revenue from EWH energy consumption and the second terms are the costs of the utility for purchasing power from the electricity market. The difference in revenue earnings for the utility ($\Delta\Omega$) between the controlled and uncontrolled EWH operations for a simulation is given as

$$\Delta\Omega = \Omega^{U(M)} - \Omega^{U}. \tag{4.9}$$

Results from (4.9) are further extrapolated to obtain the monthly $\Delta\Omega$, which are presented in Table 4.5 for winter and summer test simulations, along with the resulting yearly $\Delta\Omega$. As before, the yearly $\Delta\Omega$ is calculated assuming that winter and summer periods last for six months each.

Table 4.5: Summary of Changes to Utility Benefits ($\Delta\Omega$)

	$\Delta\Omega$ Winter ($/month)	$\Delta\Omega$ Summer ($/month)	Annual $\Delta\Omega$ ($)
Case 1 (21%)	-394	-200	-3,564
Case 2 (30%)	-510	-259	-4,614
Case 3 (50%)	-783	-392	-7,050

Both winter and summer scenarios resulted in reduced energy consumption during peak hours; 6% and 2.3% respectively. As previously mentioned, more expensive sources of power, like natural gas-fired generation plants, are utilized to meet the higher demand during peak hours. Reductions in consumption during peak hours also implies cost savings for the utility. However, Table 4.5 shows reduced net revenues (indicated by negative $\Delta\Omega$) for the EWH controlled population, for all cases. This can be attributed to utility's revenue reduction due to increased consumer savings from the incentives given by the utility. It is also noted that $\Delta\Omega$ are lower in the summer months because consumer savings in the summer are noticeably lower than in winter (see Table 4.4).

The NPV of the annual $\Delta\Omega$ stream (from Table 4.5), is calculated over five years and given in Table 4.6. The expected benefits to the utility from peak reduction, $E[B]$, is also given. The net benefit of the utility, for Case 2 and 3 indicate positive returns, but for Case 1, yields negative return. The negative return in Case 1 is attributed to the negative NPV of $\Delta\Omega$, which is greater than the expected benefits accrued to the utility from investment deferment ($E[B]$). As EWH penetration increases, the rise in benefits from investment deferment results in positive returns.

Table 4.6: Peak Reduction and Expected Benefit of Utility

	PRR	$E[B]$ ($)	NPV of $\Delta\Omega$ ($)	Net Benefit ($)
Case 1 (21%)	0.991	13,170	-13,471	-301
Case 2 (30%)	0.985	21,432	-17,440	3,992
Case 3 (50%)	0.976	32,287	-26,649	5,638

EWH Winter Operation Only

Even though Ontario is typically a summer peaking system, rural regions with lower reliance on natural gas consume more electricity than an average Ontario household. Places such as Kenora and Thunder Bay experience winter peaks, compared to more urban regions, like Toronto, where peak demand occurs in the summer [67]. It is hence assumed in this research that the shift towards household electrification will lead to winter peaking, similar to Quebec and the Atlantic provinces.

To this effect, in this sub-section, studies are carried out examining the specific benefits to the utility and consumers in Ontario considering EWH operation during winter months only. This is because in summer, the peak reduction potential is significantly low.

Table 4.7 is a revised version of Table 4.4, in which the consumer's savings during the six months of winter operation are presented, while in Table 4.8 the corresponding benefits accrued to the utility, and the gross social benefits, are given.

Table 4.7: Consumer Cost Savings to Individual Consumer from EWH Operation in Winter Operation

	Annual Savings ($)	Total Savings for Consumer Group ($)	Net Savings ($)
Case 1 (21%)	32	4,096	15,482
Case 2 (30%)	34	6,222	23,519
Case 3 (50%)	36	10,980	41,504

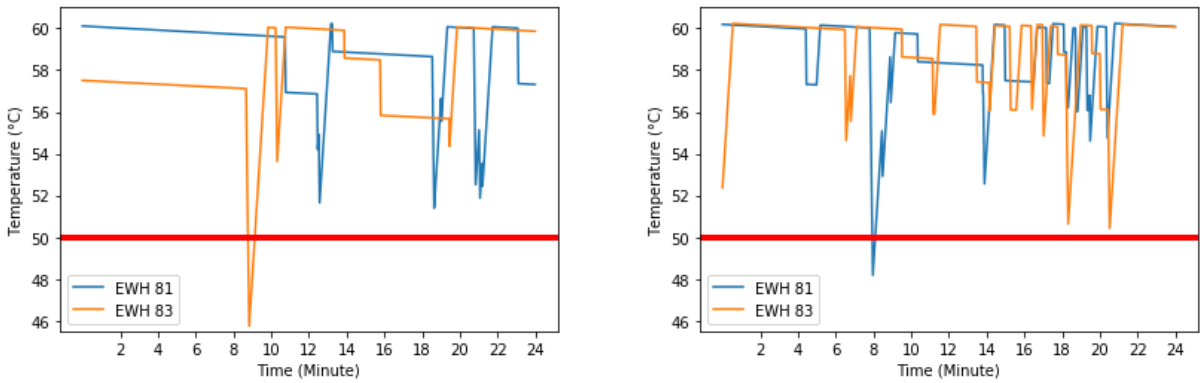
Table 4.8: Expected Utility Benefit in Winter Operation

	$E[B]$ ($)	NPV of $\Delta\Omega$ ($)	Net Benefit ($)	Social Benefit ($)
Case 1 (21%)	13,170	-8,935	4,235	19,717
Case 2 (30%)	21,432	-11,566	9,866	33,385
Case 3 (50%)	32,287	-17,758	14,529	56,033

Note that for winter operation scenario, the EWH operation is controlled for winter months only, while the summer operations remain uncontrolled. The last column in Table 4.8 provides the social benefit, which is the sum of the benefits to the utility and all consumers (sum of net utility benefit and NPV of consumer savings) for the five-year life of the EWH controller. As the EWH penetration increases, an increase in social benefit is evident as a result of the increased consumer and utility benefits. As seen in Table 4.8, using only the winter operation, the utility benefit is significantly increased, while total consumer savings is reduced due to the lack of summer savings, as seen in Table 4.7. However, the utility is now in a better position to subsidize the cost of the EWH controller, thereby providing further savings to participating consumers.

Comfort Index

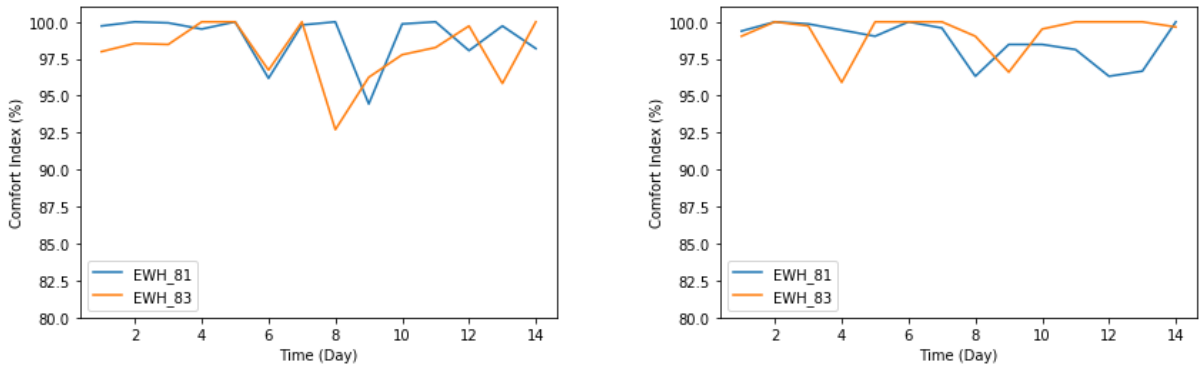
As outlined in Section 4.2.2, user comfort is defined as the percentage of time the water outlet temperature is greater than 50 °C. Figure 4.8 shows the temperature profiles of two EWHs over a given day under normal operation and proposed agent-modified operation.



(a) Normal Operation (b) Agent Modifed Operation

Figure 4.8: EWH Water Temperature

The red line marks the 50 °C threshold used in determining user comfort. Under normal operation, as seen in Figure 4.8(a), the temperature of EWH *83* dips below 50 °C for approximately 19 minutes. Under agent-modified operation, as seen in Figure 4.8(b), the temperature of EWH *81* dips below 50 °C for approximately 8 minutes. The comfort index for an entire simulation for two EWHs under normal and agent-modified operation is shown in Figure 4.9. Similar comfort index results under agent operation is noted in the other Canadian provinces examined in this thesis.



(a) Normal Operation (b) Agent Modified Operation

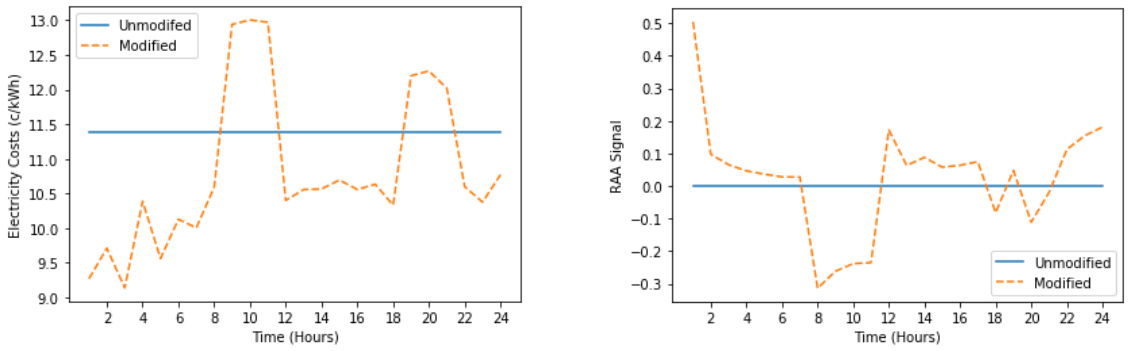
Figure 4.9: Comfort Index for a Two Week Simulation

The comfort index, as seen in Figure 4.9, is lowest on day 8 for EWH *83* under normal operation, at approximately 93% but the majority of days for the two EWHs had comfort

indexes above 98%. A comfort index of 93% corresponds to approximately 100 minutes in which outlet water temperature is less than 50 °C, likely caused by sporadic hot water usage patterns. The average comfort index for agent-modified and normal operation for the simulations shown in Figure 4.9 is 99.1% and 98.7%, respectively. Though agent-modified operation shows minor improvements in user comfort, the natural dead-band operation of EWHs usually ensures adequate hot water provisions. It is evident that regardless of the control methodologies used, 100% comfort cannot be assured as the water temperature is heavily impacted by usage patterns.

4.3.2 Case Study: New Brunswick & Quebec

Similar to the Ontario case study, the electricity tariff rates for New Brunswick and Quebec, used in this study, are listed in Tables 4.1 and 4.2. New Brunswick and Quebec, both have a fixed tariff rate for residential consumers. Given the absence of publicly available HEP data for Quebec, and since the consumer pricing structure of the two provinces are similar, the same trained RL models are used for Quebec and New Brunswick, and for the test simulations of Quebec, an HEP of zero is used. This implies that the utility in Quebec is assumed to only consider peak reduction potential from the proposed controlled EWH operations. Figure 4.10 depicts the results obtained from the MADDPG operation for the RAA and UA during one of the winter simulation days in New Brunswick.



(a) Original Tariff and the UA Modified Tariff

(b) RAA Action Signal

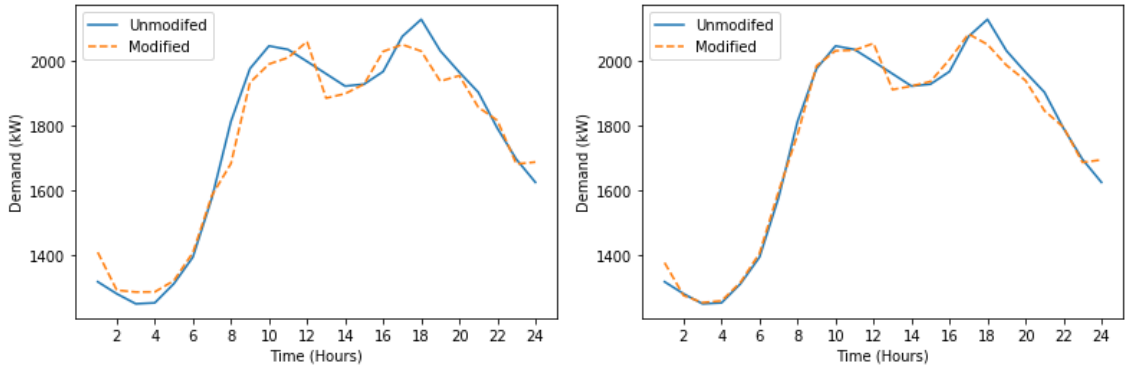
Figure 4.10: Action Signals of RAA and UA for Winter Simulations

From the UA price signal (Figure 4.10(a)), certain hours can clearly be identified as peak and non-peak hours. In this study, since the same load profile as in Ontario is

used, the peak and non-peak times are identical for the three provinces. It is important to note that the UA sends signals based on peak times (as dictated by the load profile) and its cost of electricity. The HEPs in New Brunswick and Ontario (FHMC and HOEP, respectively) are generally higher during these peak times, which leads the UA to reduce EWH consumption by increasing the residential electricity rate.

The RAA handles the activation/deactivation of a few selected EWH heating elements by sending hourly charge or discharge signals. Figure 4.10(b) illustrates the RAA action signals for the same day. Positive RAA signals in early morning and afternoon indicate the RAA's desire to increase the TES in the VB before the morning and evening peaks, respectively, which is further encouraged by the reduced electricity tariff from the UA during these hours. Negative RAA signals primarily occur during peak hours to discharge the VB.

The resulting load profiles from the multi-agent operation following the MADDPG algorithm in New Brunswick and Quebec are presented in Figure 4.11.



(a) New Brunswick (b) Quebec

Figure 4.11: Load Profile with EWH Operation

Extending the simulation over winter months, it is noted that the energy consumption during peak hours decreased by approximately 5%. This decrease is compensated by some increase in energy consumption during non-peak hours, some reduction in standby heat losses, and uncertainties in water draw profiles. Figure 4.11(a) depicts some sporadic behaviour during the evening peak because of the varying UA and RAA signals. In addition to reducing system peak, the UA aims to maximize its net revenue, which explains the noticeable decrease in demand during the peak hours of 18-21 when the HEPs are relatively

high. Similar to the Ontario scenarios, operation in summer also resulted in a decrease in energy consumption during peak hours through load shifting (approximately 1.5%).

The consumer savings from the proposed MADDPG implementation and agents' operation are presented in Table 4.9. The savings are split equally amongst participating EWHs. The average winter and summer savings from the test simulations and the annual savings obtained by extrapolating the average savings considering six months of winter and summer, are presented.

Table 4.9: Cost Savings to Individual Consumer from EWH Operation

	Winter ($/day)	Summer ($/day)	Annual Savings ($)
New Brunswick	0.168	0.069	43
Quebec	0.166	0.090	46

Using (4.6), the NPV of the consumer's savings over five years are $164 and $175 for New Brunswick and Quebec, respectively. Assuming that the cost of the EWH controller to be $100, the resulting NPV of the savings are $64 and $75, respectively.

The monthly differences in utility's revenue ($\Delta\Omega$) between the controlled and uncontrolled EWH operation, are presented in Table 4.10 for New Brunswick, for winter and summer test simulations, along with the resulting annual difference in revenue. Due to the unavailability of HEP for Quebec, its net revenues are not calculated. The FHMC from NB Power is used to calculate the cost of purchasing power for a New Brunswick utility.

Table 4.10: Summary of Changes to Utility Benefits ($\Delta\Omega$)

	$\Delta\Omega$ Winter ($/month)	$\Delta\Omega$ Summer ($/month)	Annual $\Delta\Omega$ ($)
New Brunswick	-908	-558	-8,796
Quebec	NA	NA	NA

Similar to the results obtained for Ontario (see Table 4.5), it is noted that the New Brunswick utility also incurs a reduced net revenue (indicated by negative $\Delta\Omega$) for the EWH controlled population. This revenue reduction can be attributed to increased incentives being paid by the utility to the consumers.

The NPV of the annual $\Delta\Omega$ stream (from Table 4.10) is calculated over five years and given in Table 4.11. The expected benefits to the utility from peak reduction, $E[B]$, is

also given. The net benefit to the utility in New Brunswick and Quebec indicate positive returns. It may be noted that the net benefit for the utility in Quebec is significantly higher and does not present the complete information, because the net revenue of the utility was not included due to the unavailability of HEP data, as stated earlier.

Table 4.11: Peak Reduction and Expected Benefit of Utility

	PRR	$E[B]$ ($)	NPV of $\Delta\Omega$ ($)	Net Benefit ($)
New Brunswick	0.972	37,564	-33,249	4,315
Quebec	0.982	24,725	NA	24,725

EWH Winter Operation Only

Due to the relatively low consumer savings in the summer months, and significantly lower peak reduction potential, the results assuming only winter simulation operation are considered and presented in Tables 4.12 and 4.13. The social benefit, as previously defined for Ontario, is also provided in Table 4.13.

Table 4.12: Consumer Cost Savings to Individual Consumer from EWH Operation in Winter Operation

	Annual Savings ($)	Total Savings for Consumer Group ($)	NPV of Savings ($)
New Brunswick	31	13,609	51,442
Quebec	30	8,160	30,845

Table 4.13: Expected Utility Benefit under Winter Simulation

	$E[B]$ ($)	NPV of $\Delta\Omega$ ($)	Net Benefit ($)	Social Benefit ($)
New Brunswick	37,564	-20,647	16,917	68,359
Quebec	24,725	NA	24,725	55,570

As seen in Table 4.13, considering only the winter operation, the utility benefit is significantly increased, while total consumer savings is reduced due to the absence of summer savings. However, the utility is now in a better position to subsidize the cost of the EWH controller, thereby providing further savings to participating consumers.

4.4 Discussion of Results

The proposed MADDPG implementation was evaluated on the Canadian provinces of Ontario, New Brunswick and Quebec to identify benefits to both consumers and utility. The most notable observation based on the presented results was the improved performance during the winter versus summer months. Consumer savings from all scenarios in the three provinces were nearly twice the amount in winter than in summer. This was due to the relatively coincident peaks of EWH and grid demand in winter months. Another key observation was the increased consumer savings in Ontario as compared to New Brunswick and Quebec. This is primarily due to the TOU tariff employed in Ontario, compared to the constant tariff rates in New Brunswick and Quebec. Consumer cost savings were accrued through load shifting because of EWH energy consumption reduction during the winter peak times by over 5% in all the provinces; most of the reduced load was shifted to non-peak times. Similarly, minor reductions in energy consumption for all provinces during the summer peak times was also observed. Consumers however, did not encounter any loss of comfort from lack of hot water, as the comfort index for simulations under agent operations was always above 99%, that is slightly higher than the comfort index attained during normal EWH operations. Though the comfort index was presented only for Ontario, in Section 4.3.1, similar performance for the other two provinces were also noted.

It was noted that the potential for peak reduction was greater in the winter, also attributed to the coinciding EWH and grid peaks. Increased EWH penetrations in Ontario also resulted in higher peak reduction, thereby contributing to increased benefits associated with deferment of investment. In New Brunswick and Quebec, similar peak reduction performance results were observed. However, agent performances during summer were less effective in attaining peak reduction. In addition to cost savings associated with investment deferrals, regions like New Brunswick with a greater reliance on fossil fuels for electricity generation, will benefit from peak shaving. Peak shaving can potentially reduce the need for fossil fuel generation sources in meeting peak demand, thereby reducing operation costs and GHG emissions.

Lastly, results considering only winter DSM operations were obtained and presented.

The findings indicated that though consumers received additional savings during summer operation, the utility benefits were significantly decreased due to the added operation costs in addition to the low peak reduction potential. Thus, if operating solely in the winter, utilities were able to retain more monetary benefits which placed them in a better position to subsidize the costs associated with EWH controllers. For instance, in winter operation for Ontario, assuming a subsidy of 25% (for a net cost of EWH controller cost being \$75), the resulting consumer savings for the three scenarios increased to \$46, \$54 and \$61, respectively. The resulting benefit for the utility for the three penetration scenarios were \$1,035, \$5,291 and \$6,904.

In the coming years, when EWHs are more commonly used as flexible loads and controllers are installed at the manufacturing stage, controller prices will significantly drop, resulting in further increased savings for consumers and the utility. Further studies can be carried out to assess optimal benefits for consumers and utility, considering EWH controller cost subsidies for different Canadian provinces, and for multiple residential feeders (as opposed to a single feeder considered for this work).

4.5 Summary

In this chapter, the RL models utilizing the MADDPG algorithm, proposed and developed in Chapter 3 were applied to study and assess the peak reduction and consumer cost savings potential in Ontario, New Brunswick and Quebec. First, the price and load data used in this study for simulating agent behavior was explained. Next, the provincial scenarios and various EWH penetration cases were presented. Lastly, the agent simulations were conducted and benefits accrued by consumers and utilities through agent operation under the aforementioned scenarios following the proposed methodologies were presented.

Chapter 5

Conclusion

5.1 Summary

The research presented in this thesis focused on the development of RL agents to provide energy management in a residential distribution system using EWHs as flexible loads. To this effect, the MADDPG algorithm was implemented, in which the two agents, the RAA and the UA, attained their respective objectives.

In Chapter 1, the motivations behind this research were presented and a literature review was carried out detailing the works related to EWH applications in grid services and DSM using AI. Based on the prior work and their limitations (namely, lack of scalability and generalizability), the research objectives for this thesis were established.

In Chapter 2, the main concepts required to design the RL agents were presented. The theoretical background required for generating EWH hot water draw profiles and electric load profiles was reviewed. Lastly, the MADDPG algorithm and relevant concepts related to DRL and MARL were explained in brief.

Chapter 3 proposed the agent models for the RAA and UA were introduced. A novel RAA control algorithm through a binning process was proposed and the RAA and UA state, action and reward signals were explained in detail. Lastly, the MADDPG operational interactions between the RAA and UA were explained, and the network architectures of the actor, critic, target actor and target critic networks of the RAA and UA agents were presented.

In Chapter 4, the proposed EWH control operation using the MADDPG algorithm was tested on various region-specific consumer tariffs to evaluate the impact of using EWHs as

flexible loads to attain savings for the utility and consumers. The Canadian provinces of Ontario, New Brunswick and Quebec were considered for the studies. The results indicated consumer cost savings during the winter and summer seasons, but peak reduction potential was the highest during winter. It was noted that Ontario consumers obtained slightly higher savings due to the province's TOU tariff, as compared to the savings obtained with constant tariff rates in New Brunswick and Quebec. Additional scenarios detailing winter-only operation for the three provinces were also presented and discussed. However, limitations in the analysis presented in this research such as initial assumptions (*i.e.* draw profiles) and lack of information (*i.e.* HEP for Quebec) does not allow for the full assessment of savings.

5.2 Contributions

The main contributions of the research presented in this thesis can be summarized as follows:

- The thesis presented the modelling and application of two RL agents, namely, the RAA and UA, using the MADDPG algorithm in a data-driven energy management problem for EWHs. The agents models were formulated through the creation of state, action and reward signals. Both agents interacted with each other and the environment in a manner which maximized their respective reward.
- A novel control algorithm using a binning process, to be used by the RAA, was proposed to leverage the flexibility potential of EWHs to provide monetary benefits to the consumer. This was attained via an intelligent decision making process in which the RAA strategically turned ON or OFF the EWHs based on real-time water temperature. The proposed control algorithm ensured that the comfort and availability of hot water for residential households was not compromised. The UA in turn sought to reduce its operational costs and peak demand by leveraging the RAA operation.
- The research presented herein, utilized MARL theory and implemented the MADDPG algorithm to govern the interaction between the RAA and UA. This was achieved through the actor-critic architecture in which the NNs for the actor, critic, target actor and target critic were used in guiding the RL training and testing process.

5.3 Future Work

Based on the work presented in this thesis, potential topics for future research are presented below:

- In this study, only EWHs with tank capacity of 270 L were considered. Future research could investigate the grid impact when using EWHs of multiple sizes, even EWHs in commercial and industrial sectors. Similarly, future research could investigate potential grid impact for EWHs with multiple heating elements.
- This study focused solely on energy management of the EWHs. Further work could include the implementation of multiple agents in a larger network and address additional constraints such as bus voltage and feeder current limits, and other objectives.
- In addition to the generalizability of the MARL approach, the presented studies demonstrated the scalability of the RL approach. For instance, in the proposed framework, two agents- the RAA, which controlled the EWH behaviour and the UA which emulated the utility, were considered. Additional agents could be included in the future to further establish control of other aspects of the problem. To this effect, Figure 5.1 presents a possible future concept of the RL problem within a residential feeder.

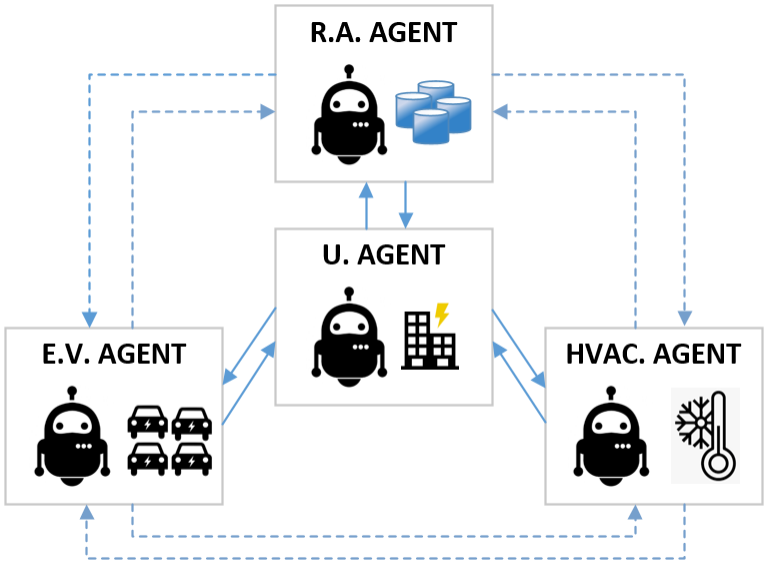


Figure 5.1: Additional Agents in a Residential Feeder Network

The above figure includes agents for EV and HVAC control. The inclusion of additional agents will require more computational power and hence a computer equipped with GPU processor is highly recommended for such development and implementation. Similarly, the impact of additional agents on the overall network need to be evaluated to ensure adequate performance.

- Using RL, applications can extend far beyond the residential sector and can aid in providing control services to the entire distribution network. Additional agents can be used to model urban neighbourhoods, commercial complexes, large industrial buildings, and wind/solar farms, to name a few. The benefits of data-driven and model-free RL, permits applications in complex systems. Future work could involve adding relevant RL agents to incrementally build the electricity network.

References

[1] Hydro Quebec. Breakdown of a household's electricity use, 2021. https://www.hydroquebec.com/residential/customer-space/electricity-use/electricity-consumption-by-use.html.

[2] K. Mason and S. Grijvala. A review of reinforcement learning for autonomous building energy management. *Computers & Electrical Engineering, Volume 78*, 2019.

[3] The Population Division of the Department of Economic and Social Affairs. World urbanization prospects: The 2014 revision. Report, United Nations, 2014.

[4] I. Dusparic, A. Taylor, A. Marinescu, V Cahill, and S. Clarke. Maximizing renewable energy use with decentralized residential demand response. In *2015 IEEE First International Smart Cities Conference (ISC2)*. IEEE, 2015.

[5] Natural Resources Canada Canada.ca. Energy and greenhouse gas emissions (GHGs), 2018. https://www.nrcan.gc.ca/science-data/data-analysis/energy-data-analysis/energy-facts/energy-and-greenhouse-gas-emissions-ghgs/20063#L6.

[6] Natural Resources Canada Canada.ca. Renewable energy facts, 2019. https://www.nrcan.gc.ca/science-data/data-analysis/energy-data-analysis/energy-facts/renewable-energy-facts/20069#L7.

[7] J. Hanania, K. Stenhouse, and J. Donev. *Intermittent electricity - Chapter 4*. University of Calgary. https://energyeducation.ca/encyclopedia/Intermittent_electricity#:~:text=Sources%20of%20intermittent%20electricity%20include,can%20be%20created%20is%20limited.

[8] B.R. Alamri and A.R. Alamri. Technical review of energy storage technologies when integrated with intermittent renewable energy. In *International Conference on Sustainable Power Generation and Supply*. IEEE, 2009.

[9] Office of Electricity U.S Department of Energy. Renewable energy integration. https://www.energy.gov/oe/services/technology-development/renewable-energy-integration.

[10] M. Oprisan and S. Pneumaticos. Potential for electricity generation from emerging renewable sources in canada. In *IEEE EIC Climate Change Conference*. IEEE, 2006.

[11] CanWEA - Canadian Wind Energy Association. Installed capacity. https://canwea.ca/wind-energy/installed-capacity/.

[12] C. Goldenberg and M. Dyson. Demand flexibility the key to enabling a low-cost, low-carbon grid. Report, Rocky Mountain Institute, 2018. https://rmi.org/wp-content/uploads/2018/02/Insight_Brief_Demand_Flexibility_2018.pdf.

[13] C. Aguilar, D.J. White, and D Ryan. Domestic water heating and water heater energy consumption in canada. Report, Canadian Building Energy End-Use Data and Analysis Centre, 2005.

[14] Natural Resources Canada Office of Energy Efficiency. Canada's secondary energy use (final demand) by sector, end use and subsector, 2018. https://oee.nrcan.gc.ca/corporate/statistics/neud/dpa/showTable.cfm?type=HB&sector=aaa&juris=ca&rn=2&page=0#sources.

[15] Natural Resources Canada Office of Energy Efficiency. Table 34: Water heater stock by building type and energy source, comprehensive energy use database, 2018. https://oee.nrcan.gc.ca/corporate/statistics/neud/dpa/menus/trends/comprehensive_tables/list.cfm.

[16] S. Dery, A. Wadhera, S. Wong, and L.P. Proulx. Real-world implementation of residential thermostat control for dr. In *2018 IEEE Canadian Conference on Electrical & Computer Engineering (CCECE)*. IEEE, 2018.

[17] M.S. Ibrahim, W. Dong, and Q. Yang. Machine learning driven smart electric power systems: Current trends and new perspectives. *Applied Energy, Volume 272, 15 August 2020*, 2020.

[18] Y. Chen, Y. Tan, and D. Deka. Is machine learning in power systems vulnerable? In *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*. IEEE, 2018.

[19] A. Moreau. Control strategy for domestic water heaters during peak periods and its impact on the demand for electricity. *Energy Procedia*, 2011.

[20] D. Podorson. Grid interactive water heaters - how water heaters have evolved into a grid scale energy storage medium. In *ACEEE Summer Study on Energy Efficiency in Buildings*. ACEEE, 2016.

[21] S. Kundu, J. Hansen, J. Lian, and K. Kalsi. Assessment of optimal flexibility in ensemble of frequency responsive loads. In *IEEE International Conference on Smart Grid Communications (SmartGridComm)*. IEEE, 2017.

[22] R. Diao, S. Lu, M. Elizondo, E. Mayhorn, Y. Zhang, and N. Samaan. Electric water heater modeling and control strategies for demand response. In *IEEE Power and Energy Society General Meeting*. IEEE, 2012.

[23] Z. Xu, R. Diao, S. Lu, J. Lian, and Y. Zhang. Modeling of electric water heaters for demand response: A baseline pde model. *IEEE Transactions on Smart Grid (Volume: 5, Issue: 5, Sept. 2014)*, 2014.

[24] A. Belov, V. Kartak, A. Vasenev, N. Meratnia, and P.J.M. Havinga. A hierarchical scheme for balancing user comfort and electricity consumption of tank water heaters. In *IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*. IEEE, 2016.

[25] T. Borsche, F. Oldewurtel, and G. Andersson. Scenario-based mpc for energy schedule compliance with demand response. In *The International Federation of Automatic Control Volume 47, Issue 3, 2014, Pages 10299-10304*. Elsevier Ltd, 2014.

[26] M.T. Ahmed, P. Faria, and Z. Vale. Financial benefit analysis of an electric water heater with direct load control in demand response. In *IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*. IEEE, 2018.

[27] T. Clarke, T. Slay, C. Eustis, and R.B. Bass. Aggregation of residential water heaters for peak shifting and frequency response services. *IEEE Open Access Journal of Power and Energy (Volume: 7)*, 2019.

[28] Y.M. Atwa, E.F. El-Saadany, and M.M. Salama. Dsm approach for water heater control strategy utilizing elman neural network. In *IEEE Canada Electrical Power Conference*. IEEE, 2007.

[29] T.A. Nakabi and P. Toivanen. An ann-based model for learning individual customer behavior in response to electricity prices. *Sustainable Energy, Grids and Networks*, 2018.

[30] I. Antonopoulos, V. Robua, B. Couraud, D. Kirli, S. Norbu, A. Kiprakis, S. Elizondo-Gonzalez, and S. Wattam. Artificial intelligence and machine learning approaches to energy demand-side response: A systematic review. *Renewable and Sustainable Energy Reviews Volume 130, September 2020, 109899*, 2020.

[31] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2017.

[32] C. Patyn, F. Ruelens, and G. Deconinck. Comparing neural architectures for demand response through model-free reinforcement learning for heat pump control. In *IEEE International Energy Conference (ENERGYCON)*. IEEE, 2018.

[33] F. Ruelens, J. Claessens, S. Quaiyum, R. Babuska, and R. Belmans. Reinforcement learning applied to an electric water heater: From theory to practices. *IEEE Transactions on Smart Grid (Volume: 9, Issue: 4, July 2018)*, 2018.

[34] A. Sheikhi, M. Rayati, and A.M. Ranjbar. Dynamic load management for a residential customer; reinforcement learning approach. *Sustainable Cities and Society*, 2015.

[35] S. Rozada, D. Apostolopoulou, and E. Alonso. Load frequency control: A deep multi-agent reinforcement learning approach. In *2020 IEEE Power & Energy Society General Meeting (PESGM)*, 2020.

[36] L. Hurtado, E. Mocanu, P. Nguyen, M. Gibescu, and R. Kamphuis. Enabling cooperative behavior for building demand response based on extended joint action learning. *IEEE Transactions On Industrial Informatics, Volume. 14, NO. 1, January 2018*, 2018.

[37] F. Golpayegani, I. Dusparic, A. Taylor, and S. Clarke. Multi-agent collaboration for conflict management in residential demand response. *Computer Communications Volume 96, 15 December 2016, Pages 63-72*, 2016.

[38] X. Xu, Y. Xu, S. Xu, C.S. Lai, Y. Jia, and S. Chai. A multi-agent reinforcement learning-based data-driven method for home energy management. *IEEE Transactions on Smart Grid, February 2020*, 2020.

[39] S. Wong, W. Muneer, S. Nazir, and A. Prieur. Designing, operating, and simulating electric water heater populations for the smart grid. Report, CanmetENERGY, 2013.

[40] Hydro Solution. What is the best temperature for your water heater? https://www.hydrosolution.com/en/guide-and-tips/what-is-the-best-temperature-for-your-water-heater/.

[41] Office of Energy Saver U.S Department of Energy. Storage water heaters. https://www.energy.gov/energysaver/water-heating/storage-water-heaters.

[42] M. Glavic, R. Fonteneau, and D. Ernst. Reinforcement learning for electric power system decision and control: Past considerations and perspectives. *IFAC*, 2017.

[43] S. Zychlinski. The complete reinforcement learning dictionary, 2019. https://towardsdatascience.com/the-complete-reinforcement-learning-dictionary-e16230b7d24e#:~:text=Episode%3A%20All%20states%20that%20come,we%20consider%20an%20infinite%20episode.

[44] B. Ding, H. Qian, and J. Zhou. Activation functions and their characteristics in deep neural networks. In *2018 Chinese Control And Decision Conference (CCDC)*. IEEE, 2018.

[45] H. Lee, Y. Aizawa, and K. Abe. Reinforcement learning for continuous state spaces based on locally weighted regression. In *SICE AnnuaI Conference in Sapporo, August 4-6, 2004*. IEEE, 2004.

[46] C. Nicholson. A beginner's guide to deep reinforcement learning. https://wiki.pathmind.com/deep-reinforcement-learning.

[47] T. Lillicrap and J. Hunt. Continuous control with deep reinforcement learning, 2016.

[48] R. Sutton, D McAllester, S Singh, and Y Mansour. Policy gradient methods for reinforcement learning with function approximation, 2000.

[49] Keras.io Github. Deep deterministic policy gradient (ddpg), 2020. https://keras.io/examples/rl/ddpg_pendulum/#:~:text=Deep%20Deterministic%20Policy%20Gradient%20(DDPG)%20is%20a%20model%2Dfree,algorithm%20for%20learning%20continous%20actions.&text=It%20uses%20Experience%20Replay%20and,operate%20over%20continuous%20action%20spaces.

[50] Quadrennial Technology Review. Chapter 3: Enabling modernization of the electric power system. flexible and distributed energy resources. Report, U.S Department of Energy, 2015.

[51] M. Bowling and M. Veloso. An analysis of stochastic game theory for multiagent reinforcement learning. 2000.

[52] S. O'Connell and S. Riverso. Flexibility analysis for smart grid demand side services incorporating 2nd life ev batteries. In *2016 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, 2016.

[53] A. Bin Humayd and K. Bhattacharya. Design of optimal incentives for smart charging considering utility-customer interactions and distribution systems impact. *IEEE Transactions on Smart Grid (Volume: 10, Issue: 2, March 2019)*, 2019.

[54] J. Flynt. 6 best smart water heaters and controllers of 2019, 2019. https://3dinsider.com/smart-water-heaters/.

[55] The Independent Electricity System Operator (2016). Hourly ontario energy price (hoep). http://reports.ieso.ca/public/PriceHOEPPredispOR/.

[56] Energie NB Power (2016). Final hourly marginal cost. https://tso.nbpower.com/Public/en/op/market/report_list.aspx?path=.

[57] Ontario Energy Board. Managing costs with time-of-use rates. https://www.oeb.ca/rates-and-your-bill/electricity-rates/historical-electricity-rates.

[58] Hydro Quebec. Rate for residential and farm customers. https://www.hydroquebec.com/residential/customer-space/rates/rate-d.html.

[59] Energie NB Power (2016). Residential rates. https://www.nbpower.com/en/products-services/residential/rates.

[60] Energyrates.ca. Québec Electricity and Natural Gas Options. https://energyrates.ca/quebec/.

[61] Data and Canada Energy Regulator Analysis. Canada's renewable power landscape 2017 – energy market analysis, 2017. https://www.cer-rec.gc.ca/en/data-analysis/energy-commodities/electricity/report/2017-canadian-renewable-power/province/canadas-renewable-power-landscape-2017-energy-market-analysis-new-brunswick.html.

[62] J. Erickson. *Electric Distribution Manual*. SDG&E, San Diego, CA, USA, June 2020. [Online]. https://www.sdge.com/sites/default/files/DM_0.pdf.

[63] J. Leadbetter and L. Swan. Battery storage system for residential electricity peak demand shaving. *Energy and Buildings Volume 55, December 2012, Pages 685-692*, 2012.

[64] J. Eyer. Electric utility transmission and distribution upgrade deferral benefits from modular electricity storage. Report, Sandia National Laboratories, 2009.

[65] V. Dufresne. The value of demand response in a hydro-dominated power grid – the example of quebec, canada, 2016.

[66] Investopedia. Calculating Present and Future Value of Annuities. https://www.investopedia.com/retirement/calculating-present-and-future-value-of-annuities/.

[67] IESO. Demand Overview - Historical Demand. https://www.ieso.ca/en/Power-Data/Demand-Overview/Historical-Demand.

APPENDICES

Appendix A

Determining the Number of Houses Connected to Residential Feeder Using Base Load Data

A.1 EWH Distribution in Ontario

This section presents the process utilized in calculating the number of houses and number EWHs in an Ontario feeder. The parameters used in the calculations are listed below and are also given in Chapter 4, and in particular, Table 4.3.

- Household Maximum Demand ($P^{HS,MAX}$) = 7.2 kW
- Maximum Feeder Load (($P^{F,MAX}$) = 2.2 MW (2200 kW)
- Diversity Factor (DF) = 0.5
- Province-wide EWH Penetration (EWH^P) = 21%
- Number of houses supplied by feeder (N^{HS}) and number of EWHs (N^{EWH})

$$N^{HS} = \frac{P^{F,MAX}}{DF \cdot P^{HS,MAX}} \tag{A.1}$$

$$N^{EWH} = N^{HS} \cdot EWH^P \tag{A.2}$$